Revealing Hidden Patterns in Political News and Social Media with Machine Learning

Konstantina Lazaridou
March 2nd 2021, PhD defense
Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam

Outline

- **Motivation for political data analysis**
- Thesis contributions
- Political media bias detection with human annotations
- Expert and non-expert news corpora
- Article classification with curriculum learning
- Bias detection results
- Summary and findings

Why using political text data?

- Practical applications: classifying, summarizing and recommending political content
- Social impact: understanding viewpoints of the media and the general public

The New York Times

Trump, in Taped Call, Pressured Georgia Official to 'Find' Votes to Overturn Election

The president vaguely warned of a "criminal offense" as he pressured Secretary of State Brad Raffensperger in the call, according to an audio recording.

6331

Comments 6331

Share your thoughts.

The Times needs your voice. We welcome your on-topic commentary, criticism and expertise. Comments are moderated for civility.

NYT Picks | Reader Picks | All

Phil Witt — Times Pick
New York City | 2h ago

Will someone please enlighten me how, notwithstanding the illegality or repugnance of this call, overturning the Georgia results will, as reported, overturn the election, making Trump the winner. What am I missing? There are 16 electoral votes allotted to Georgia. How does that, were it fathomable, reduce Biden's total electoral count, to under 270?
Can someone please explain?

29 Replies | 164 Recommend | Share | Flag

Marge Keller — Times Pick
Midwest | 4h ago

Yet again, nothing is or will be done by the Republican Senate over a second dubious conversation Trump has had to garner favors from folks for his own personal gain.

In January, this is the same Republican Senate who flatly refused

The New York Times

ines › | Vaccine Questions | Doses Per State | After Your Vaccine | How the Moderna Vaccine Works | Why You

Opinion

THE CONVERSATION

Is It Too Late to Impeach Trump Again?

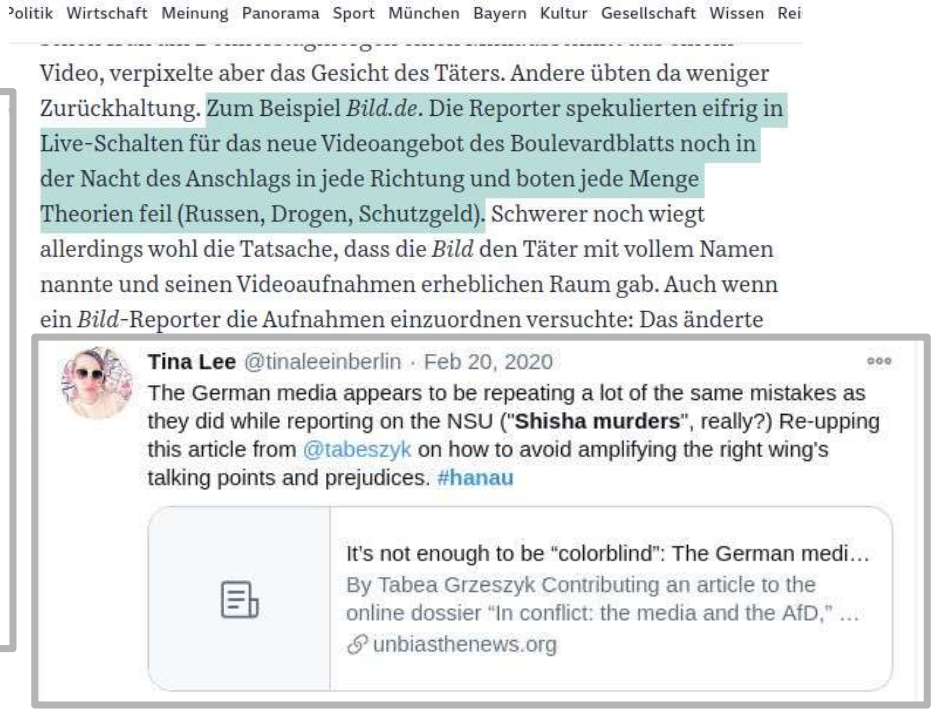
The president showed this weekend that he still has Georgia on his mind.

By Gail Collins and Bret Stephens
Ms. Collins and Mr. Stephens are opinion columnists. They converse every week.

Jan. 4, 2021

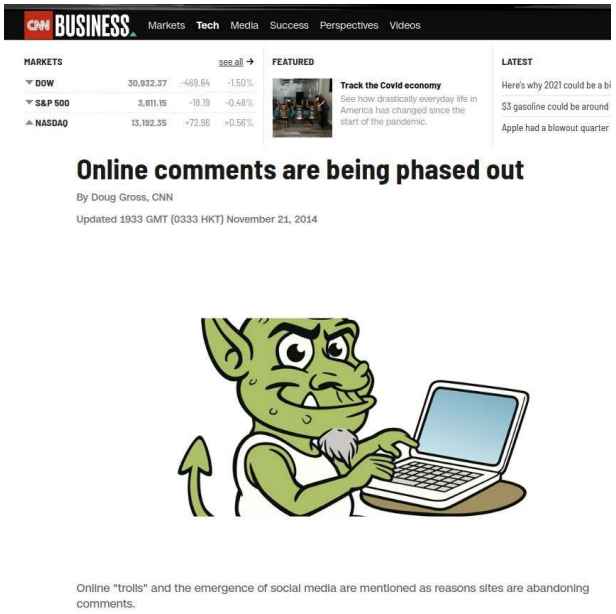
Why analyzing political media?

- Ensure a quality media ecosystem
- Promote transparent democracy
- Raise political awareness and participation



Motivation for automated solutions

- Too large and **complex data** for manual analysis
- Cumbersome to obtain large amounts of **reliable human annotations**
- Machines can iteratively find **hidden and new patterns** in the text
  - Biased **political citations** in the news
  - User **political affiliations** from reactions to news stories
  - Subtle **media bias** manifestations in news articles



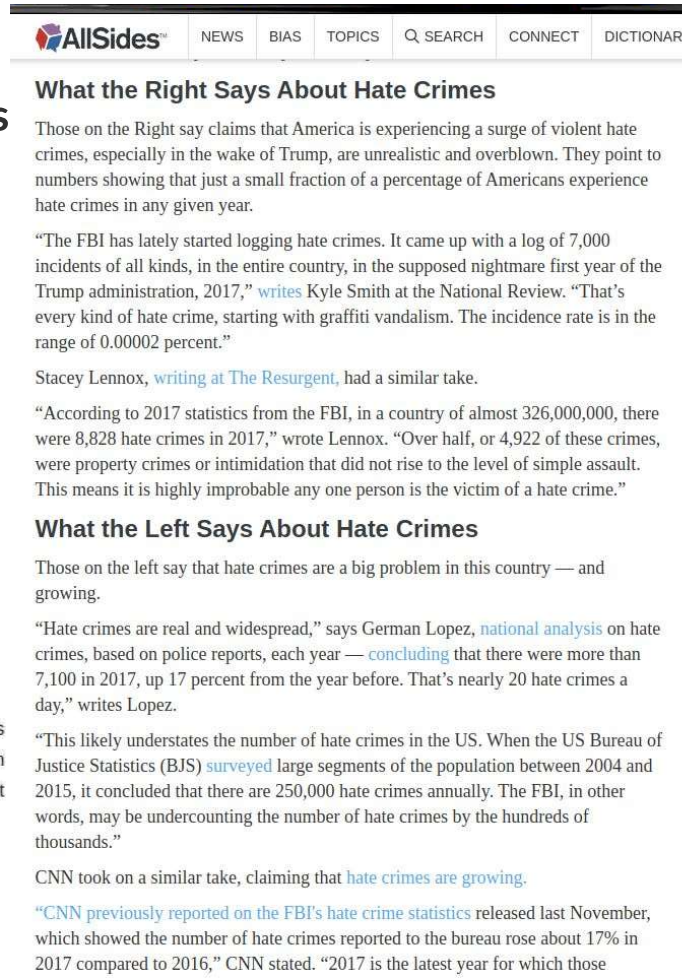
The New York Magazine

Iraq forces suspension of U.S. surveillance flights

USA Today

U.N. Withdraws U-2 Planes

It is clear that each headline reflects the same events. As far as transitivity structure is concerned here, the way participants in each clause are structured, as well as the words chosen to report the occurrence, are illustrative of how language is used differently to promote or reflect different interests.



Outline

- Motivation for political data analysis
- **Thesis contributions**
- Political media bias detection with human annotations
- Expert and non-expert news corpora
- Article classification with curriculum learning
- Bias detection results
- Summary and findings

Thesis contributions

Media Distinctiveness in Political News

Classifying reported speech
Classifying news comments

[TPDL 15, LWDA 16, ICDM 17, NLE 21*]

Political Media Bias Detection

Classifying news articles

[LREC 2020]

Political Discourse in User Generated Content

Classifying tweets

[WWW 17, WebSci 18]

*under revision

Outline

- Motivation for political data analysis
- Thesis contributions
- **Political media bias detection with human annotations** [LREC 2020]
- Expert and non-expert news corpora
- Article classification with curriculum learning
- Bias detection results
- Summary and findings

[LREC 2020] Konstantina Lazaridou, Alexander Loeser, Maria Mestre, and Felix Naumann. Discovering biased news articles leveraging multiple human annotations. In International Language Resources and Evaluation Conference. European Language Resources Association, 2020.

Problem statement

Media bias definition

A biased news article leans towards or against a certain person or opinion by making one-sided, misleading or unfair judgements.
An unbiased news article reports fair, impartial and objective information.

Problem definition

Given a news article, classify it into politically biased or unbiased.

Research questions

Is domain expertise necessary for this task?

Can deep learning techniques solve this task?

Limitations of previous work

- Non scalable (semi-) **manual studies** in political science
  [Hamborg et al. 2018., Groseclose et al. 2005, Gentzkow et al. 2010]
- Computer science studies that focus on **opinionated texts** or **political perspective identification**
  [Cohen et al. 2013, Iyyer et al. 2014, Recasens et al. 2013, Aires et al. 2019, Niculae et al. 2015, IR Journal 2018, Saleh et al. 2019]
- **Unreliable data annotations** due to commonly known taxonomies
  [Umarova et al. 2019, Aires et al. 2019, Niculae et al. 2015, Saleh at al. 2019]
- **Approximating** the bias detection problem
  [Niculae et al. 2015, Zhou et al. 2011, Kulshrestha et al. 2018, Chen et al. 2018]
- **Ungeneralizable feature** engineering
  [Volkova et al. 2018, Saez-Trumper et al. 2013]

Outline

- Motivation for political data analysis
- Thesis contributions
- Political media bias detection with human annotations
- **Expert and non-expert news corpora**
- Article classification with curriculum learning
- Bias detection results
- Summary and findings

News corpora

Dataset	Articles	Annotations/ article	Annotation level	Labels	News sources	Biased articles	Unbiased articles	Inter annotator agreement (ITA)[2]
Experts **(E)**	1,154	3	sentence	binary	306	523	631	0.59
Non-experts **(NE)**	2,993	3	document	1-5	961	1,197	1,230	0.66
Publishers **(P)**[1]	750,000	1	source	binary	1,194	375,000	375,000	-

1 Semeval 2019 hyper partisanship competition: https://pan.webis.de/semeval19/semeval19-web/

2 Krippendorff's alpha: https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

Data transformation & cross-dataset comparison

- Humanly labeled datasets (E, NE): Aggregated human labels with Dawid Skene scheme

  [Dawid et al. 1979, Sinha 2018]

- Expert labeled dataset (E): Propagated sentence-level expert labels to the document level
- Crowd sourced dataset (NE): Binarized crowd-sourced labels with the two ends of the score range

  [Maas et al. 2011]

- Computed expert vs. non-expert annotator agreement [Snow et al. 2011]

Dataset comparison	Cross-dataset agreement	STDEV %
E vs. NE	73.68	17.13
E vs. E	65.46	15.46
E vs. All	67.39	7.87
NE vs NE	64.76	19.51
NE vs All	64.32	20.30

Outline

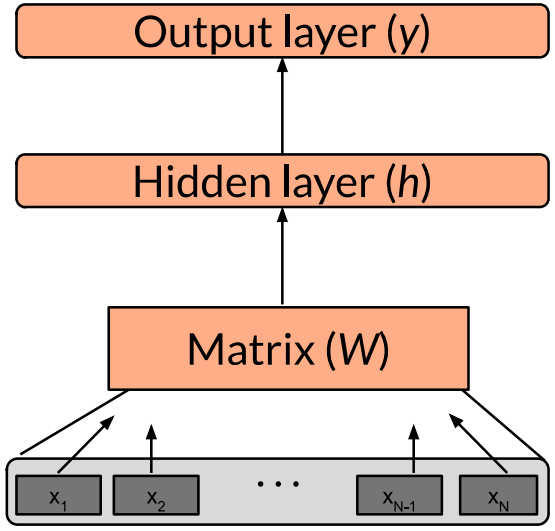
- Motivation for political data analysis
- Thesis contributions
- Political media bias detection with human annotations
- Expert and non-expert news corpora
- **Article classification with curriculum learning**
- Bias detection results
- Summary and findings

Classification methods

Baseline

fastText

hierarchical softmax



averaged bag of n-grams [Joulin et al. 2017]

Curriculum learning with humans in the loop

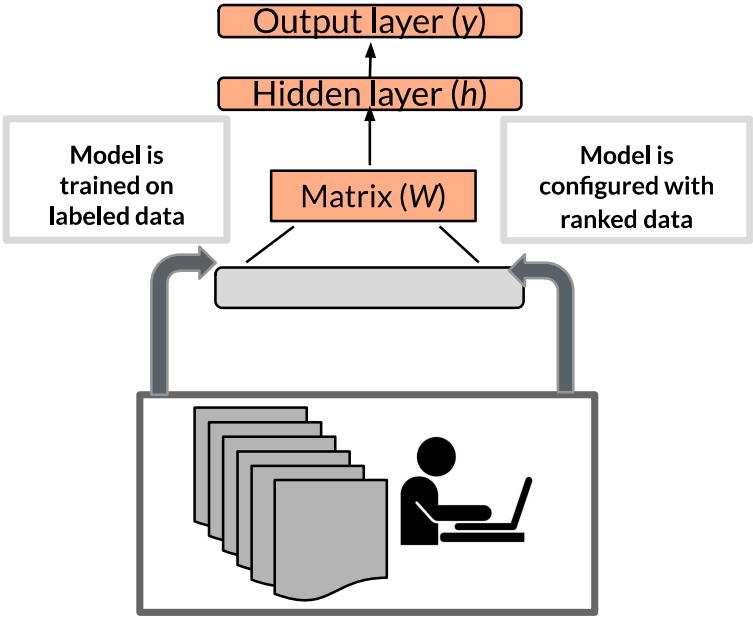
- Initiating a **curriculum** that helps the model learn
  [Bengio et al. 2009, Weinshall et al. 2018]
- Defining the **learning difficulty** of the data
- **Sorting training data** according to the learning difficulty
- **Incremental training** of the model

Our approach

Our learning difficulty

- Agreement corresponds to clearly biased or unbiased content
- "Easy" to learn article: high annotator agreement[1,2]
- "Difficult" to learn article: low annotator agreement[3,4]

- Train models with the top-10%, top-20%, …, top-100%
- Load previous weights in each iteration



1 "Obama speaks out after GOP health care bill failure", aol.com
2 "Biden does it in front of us! 'Outraged' Dems won't talk about creepy Joe's inappropriate touching", bizpacreview.com
3 "How To Call Your Representatives For Gun Control Reform", bustle.com
4 "Donald Trump slams Hillary Clinton again with "woman's card"", cbsnews.com

Outline

- Motivation for political data analysis
- Thesis contributions
- Political media bias detection with human annotations
- Expert and non-expert news corpora
- Article classification with curriculum learning
- **Bias detection results**
- Summary and findings

Test dataset

- Use **small humanly labeled dataset** (similarly to Semeval 2019)
- Articles labeled by **experts and non-experts** (E∩NE)
- Articles with **agreement on each label**
- Random sample of 40%
- **Reliable and universal** test set

Training sets			Test set
E	NE	P	⊂(E∩NE)
759	1,805	750,000	237

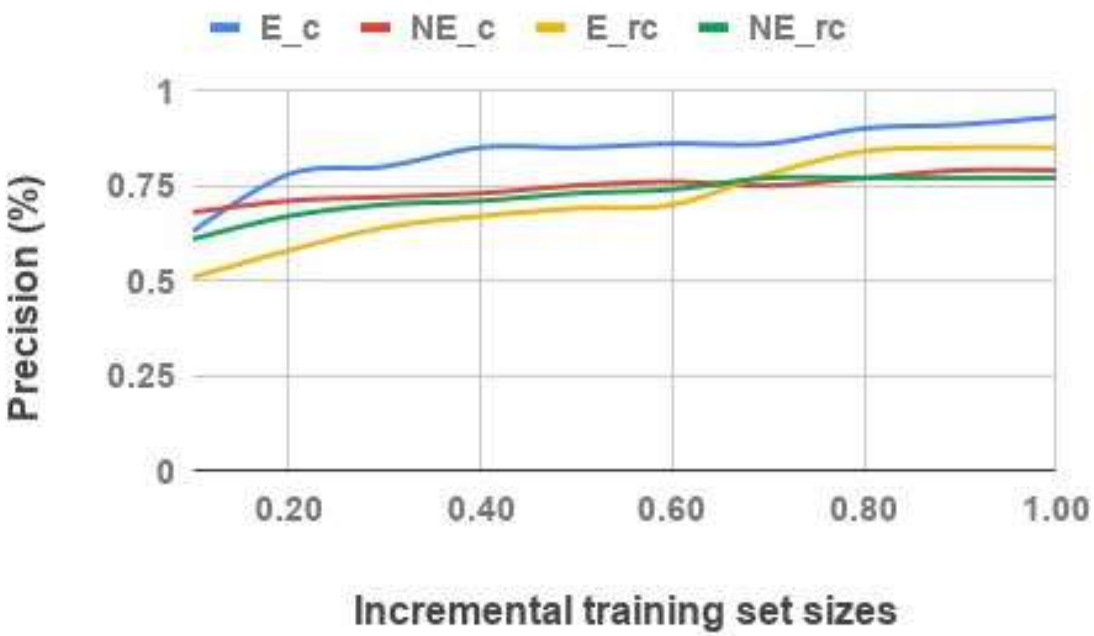
Classification results

- Domain expertise and consensus prevail over crowd sourced and publisher-based data

Training data	Precision	Recall	F-1
E	0.90	0.89	0.89
NE	0.85	0.89	0.87
E_curriculum	**0.93**	**0.95**	**0.93**
NE_curriculum	0.79	0.86	0.82
E_reverse	0.85	0.90	0.88
NE_reverse	0.78	0.87	0.83
P	0.54	0.89	0.67

Classification results

- Small amount of quality labels are more helpful than large automatically labeled data



Error analysis for the *E_c* model

- **Loaded language** for 50% of misclassifications (opinionated word by journalist or loaded topic)
  - An article calls Donald Trump “misogynist”
  - The same article is discussing the topic of “sexual assault”
- **Mixture of facts and opinions** for 60% of misclassifications (disguised columns, satire)
- **Article topic is a politician** for 60% of misclassifications (mostly false positives)
- **Sensitive topics** for 13% of misclassifications (mostly false positives)
  - An article’s headline is opinionated (“Michelle Obama slammed Donald Trump"), discusses Trump’s comments on women, and it is about two politicians[1]

1 https://www.msnbc.com/stephanie-ruhle/watch/michelle-obama-enough-is-enough-785974851664

Outline

- Motivation for political data analysis
- Thesis contributions
- Political media bias detection with human annotations
- Expert and non-expert news corpora
- Article classification with curriculum learning
- Bias detection results
- **Summary and findings**

Summary

- First study to introduce **humanly annotated news articles** for media bias detection
- First study to **compare experts with non-expert** for media bias detection
- Introducing **curriculum-based machine learning method** for classifying news articles successfully to biased or unbiased
- Using **human label consensus** as a curriculum for a machine learning model
- Manual error interpretation

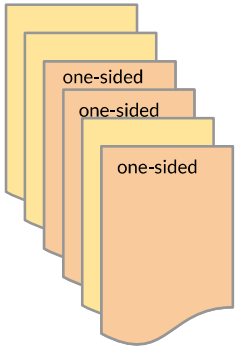
Findings

- **Human labels more suitable** than automatic labels
- Inter-annotator agreement above 60%
- Few hundreds of human labels **outperform 750,000 automatic labels**
- **Top-20% of high-consensus expert articles** achieve 80% precision
- Consensus among experts is more helpful to learner than consensus among non-experts
- Interesting error classes that can help future work
- Our proposed model E_c can be used in **research and commercial applications** to detect political media bias
- E_c can be applied for **monitoring and regulating** online content along with existing systems

Publication list

- ***Konstantina Lazaridou*** and Ralf Krestel. Identifying political bias in news articles. Special Issue of the Bulletin of the IEEE Technical Committee on Digital Libraries, 12, 2015.
- Christian Godde, Konstantina Lazaridou, and Ralf Krestel. Classification of german newspaper comments. In Proceedings of the Lernen, Wissen, Daten, Analysen Conference , volume 1670. Online Proceedings for Scientific Conferences and Workshops, 2016.
- Toni Gruetze, Ralf Krestel, Konstantina Lazaridou, and Felix Naumann. What was Hillary Clinton doing in Katy, Texas? In Proceedings of the International Conference on World Wide Web Companion. Association for Computing Machinery, 2017.
- ***Konstantina Lazaridou***, Ralf Krestel, and Felix Naumann. Identifying media bias by analyzing reported speech. In Proceedings of the IEEE International Conference on Data Mining . IEEE Computer Society, 2017.
- ***Konstantina Lazaridou***, Toni Gruetze, and Felix Naumann. Where in the world is Carmen Sandiego?: Detecting person locations via social media discussions. In Proceedings of the International Conference on Web Science. Association for Computing Machinery, 2018.
- ***Konstantina Lazaridou***, Alexander Loeser, Maria Mestre, and Felix Naumann. Discovering biased news articles leveraging multiple human annotations. In International Language Resources and Evaluation Conference. European Language Resources Association, 2020.
- ***Konstantina Lazaridou***, Alexander Loeser, Felix Naumann, and Ralf Krestel. Reported speech in political news articles: A media bias perspective. Natural Language Engineering (in revision), 2020.

Thank you for attention!



Discovering biased news articles leveraging multiple human annotations.

Is domain expertise necessary for this task?

Can deep learning techniques solve this task?

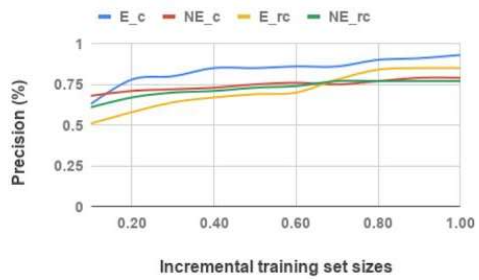
Compared sets	Agreement %	STDEV %
E vs. *NE*	73.68	17.13
E vs. *E*	65.46	15.46
E vs. *All*	67.39	7.87
NE vs. *NE*	64.76	19.51
NE vs. *All*	64.32	20.3

Training sets			Test set
E	*NE*	*P*	$\subset(E \cap NE)$
759	1,805	750,000	237

fastText

Curriculum learning with humans in the loop

Training	Precision	Recall	F-1
E_c	0.93	0.95	0.93
E_rc	0.85	0.90	0.88
NE_c	0.79	0.86	0.82
NE_rc	0.78	0.87	0.83
P	0.54	0.89	0.67



Konstantina Lazaridou

Revealing Hidden Patterns in Political News and Social Media with Machine Learning

PhD defense, Hasso Plattner Institute, March 2021

References

- Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: An interdisciplinary literature review. International Journal on Digital Libraries, 20, 2018.
- Tim Groseclose and Jeffrey Milyo. A measure of media bias. The Quarterly Journal of Economics, 120, 2005.
- Matthew Gentzkow and Jesse M Shapiro. What drives media slant? Evidence from us daily newspapers. Econometrica, 78, 2010.
- Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In Proceedings of the International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence, 2013.
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2014.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2013
- Khonzodakhon Umarova and Eni Mustafaraj. How partisanship and perceived political bias affect wikipedia entries of news sources. In Companion 138 Proceedings of the International Conference on World Wide Web. Association for Computing Machinery, 2019.
- Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. A link-based approach to detect media bias in news websites. In Companion Proceedings of the International Conference on World Wide Web. Association for Computing Machinery, 2019.
- Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-NiculescuMizil, and Jure Leskovec. Quotus: The structure of political media coverage as revealed by quoting patterns. In Proceedings of the International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015.
- Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In Proceedings of the International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence, 2011
- Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Search bias quantification: Investigating political bias in social media and web search. Information Retrieval Journal, 22, 2018.
- Tae Yano, Philip Resnik, and Noah A Smith. Shedding (a thousand points of) light on biased language. In Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, 2010.

References

- Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In Companion Proceedings of the International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018
- Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In Proceedings of the ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, 2013.
- Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Applied Statistics, Journal of the Royal Statistical Society, Series C, 28, 1979.
- Vaibhav B Sinha, Sukrut Rao, and Vineeth N Balasubramanian. Fast dawidskene: A fast vote aggregation scheme for sentiment classification. In Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining. Association for Computing Machinery, 2018
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011.
- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
- Klaus Krippendorff, Computing Krippendorff's alpha-reliability, 2011
- Emmanuel Vincent and Maria Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In Proceedings of the Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and of the Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management. Online Proceedings for Scientific Conferences and Workshops, 2018
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2017
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the ACM Annual International Conference on Machine Learning. Association for Computing Machinery, 2009

References

- Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In Proceedings of the International Conference on Machine Learning, 2018
- Tae Yano, Philip Resnik, and Noah A Smith. Shedding (a thousand points of) light on biased language. In Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, 2010
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2018
- Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In Proceedings of the International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2019
- Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Learning to flip the bias of news headlines. In Proceedings of the International Conference on Natural Language Generation. Association for Computational Linguistics, 2018.