

FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias

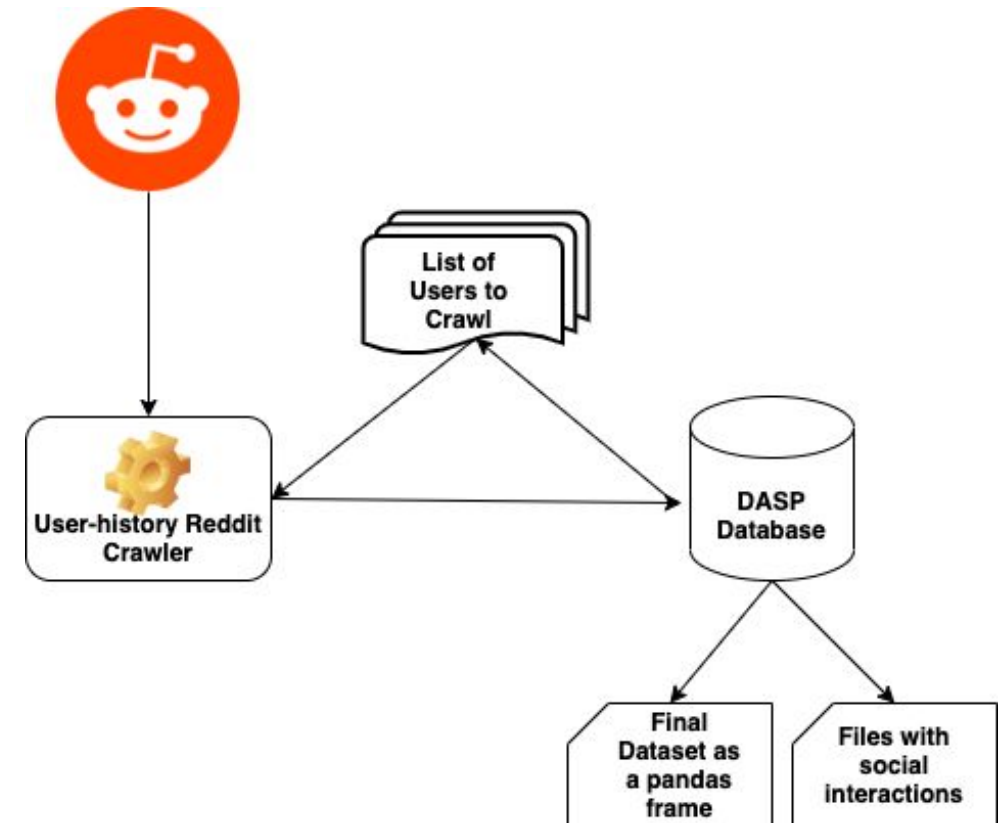
Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri-Jacques Geiss, Paolo Rosso, Lucie Flek

What is the focus of our work?

- Dissemination of online disinformation is growing
- Fact corrections frequently fail and might backfire (Redlawsk et al., 2010; Nyhan and Reifler, 2010; Swire et al., 2017; Berinsky, 2017).
- View the fake news detection task on a user level
- Explore misinformation spreading behavior together with political bias



- **Iterative** and **user-centric** approach
- Initial data from a list of 65 subreddits concerning controversial political topics we crawled 300 threads each.
- U.S. American presidential race, the SARS-CoV-2 pandemic, vaccines, abortion, feminism, gun control, climate change, 5G or politics in general



- We used **mediabiasfactcheck.com** as our source for annotated news outlet domains
- Two dimensions: **Factuality + Bias**
- Assigned after review of at least 10 headlines and 5 news stories by the site's curators
- In total 1577 misinformation domains and 571 factual domains
- This also enabled more fine-grained user-level annotation

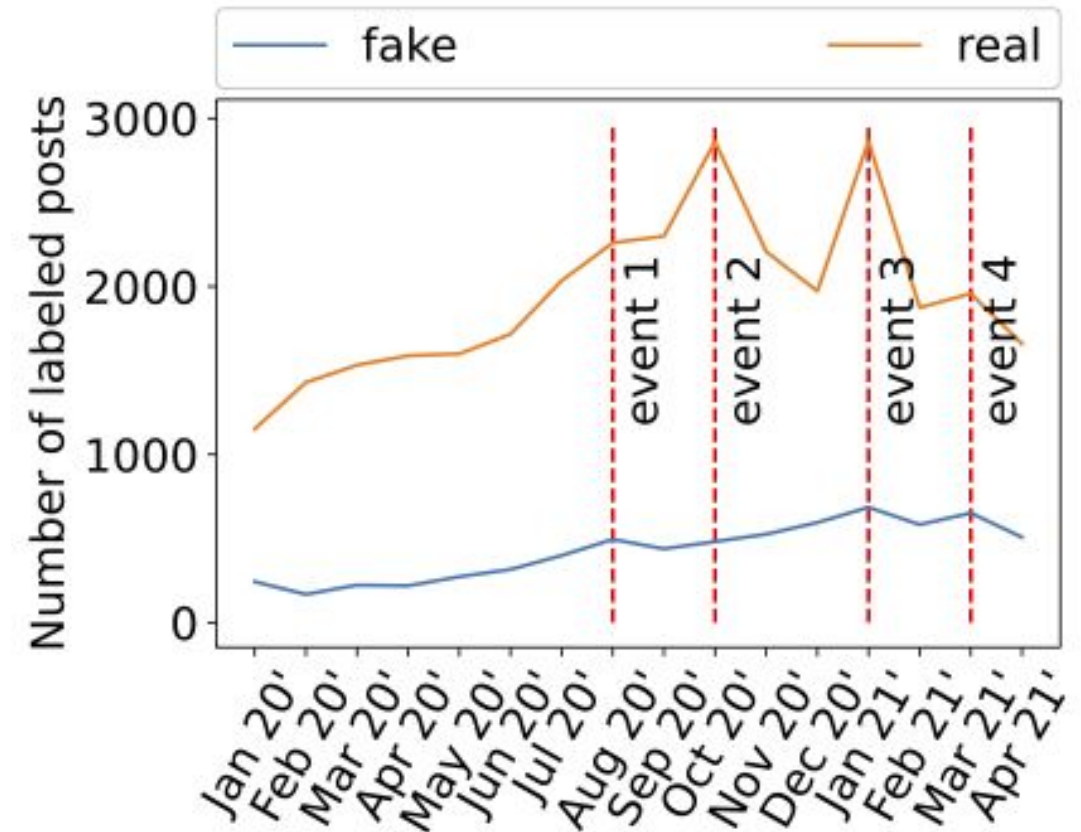


Factual Reporting
Very High
High
Mostly Factual
Mixed
LOW
Very Low

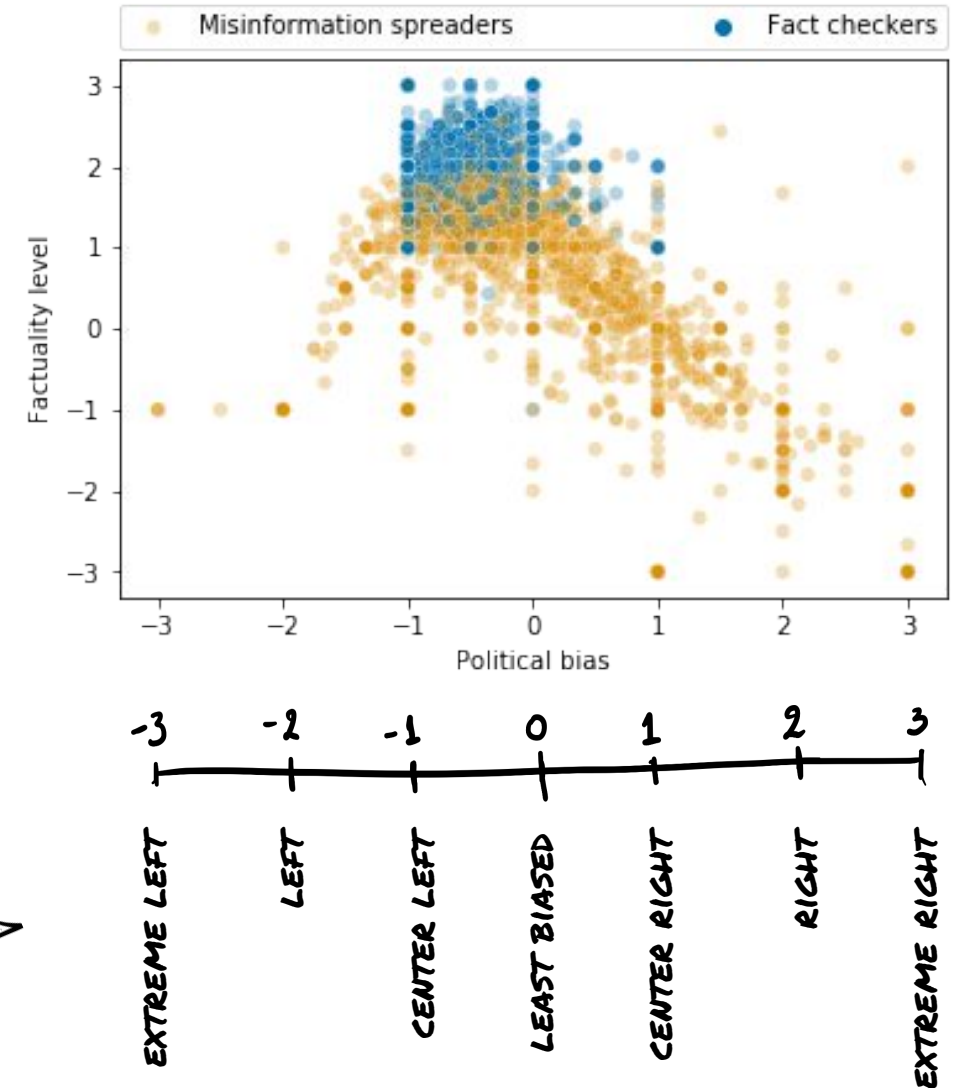
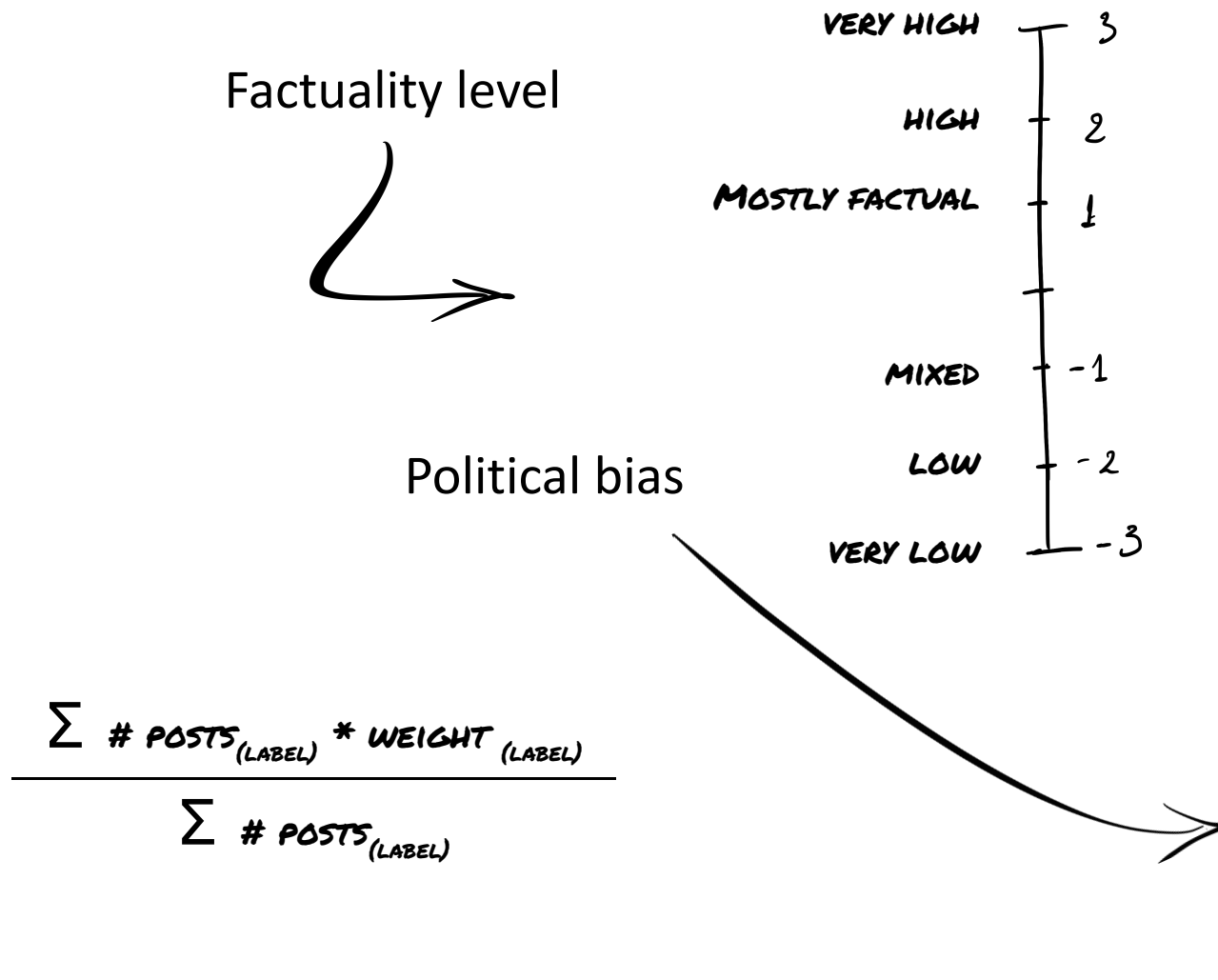
QUESTIONABLE SOURCE

News Domain Annotation

Date	Event Description
Feb 5	Trump is acquitted on the charges of abuse of power and obstruction of Congress.
Jul 11	Mail-in votes are encouraged.
Jul 30	Donald Trump threatens to postpone the election if it appears mail-in votes might go against him. (We regard this as if this had happened in August, since the effects of this political event would be still discussed during that month)
Aug 11	Joe Biden chooses Senator Kamala Harris (D-CA) as his running mate (event 1)
Nov 3	2020 United States elections (event 2)
Jan 6	US Capitol is attacked by supporters of Trump (event 3)
Feb 24	Johnson \& Johnson's vaccine candidate receives emergency use authorization from the FDA (event 4)



Political Bias and Factuality



- In total around **4150** crawled users
- **74:26** class distribution **fact-checkers** vs **misinformation spreaders**
- **2.3 million posts** from Jan 2020 until April 2021
- On average **1240** posts of **misinformation spreaders** and **654** of **fact-checkers**
- **42%** of people with **left** wing bias were **misinformation spreaders**
- **92%** of people with **right** wing bias were **misinformation spreaders**

User Embeddings

- **UBERT** - Sentence-BERT (SBERT) to obtain user representations based on post embeddings
- **User2Vec** - optimizing the conditional probability of texts given the author
- **Psycho linguistic** - using the Big Five Model and LIWC software to capture users' personality traits and mental processes

Models

- Support Vector Machine (**SVM**)
- Logistic Regression (**LogReg**)
- Random Forests (**RnFor**)
- Graph Attention Networks (**GAT**)

Fake News Spreader Detection

Model	F1 score
GAT + User2Vec (200)	61.6%
GAT + UBERT (768)	61.2%
GAT + Psycholing (83)	53.6%
GAT + User2Vec + Psycholing (283)	59.4%
GAT + Random (200)	47.8%


- Comparison of different user embeddings
- F1-score over a 5-fold CV
- User2Vec embeddings best performing one

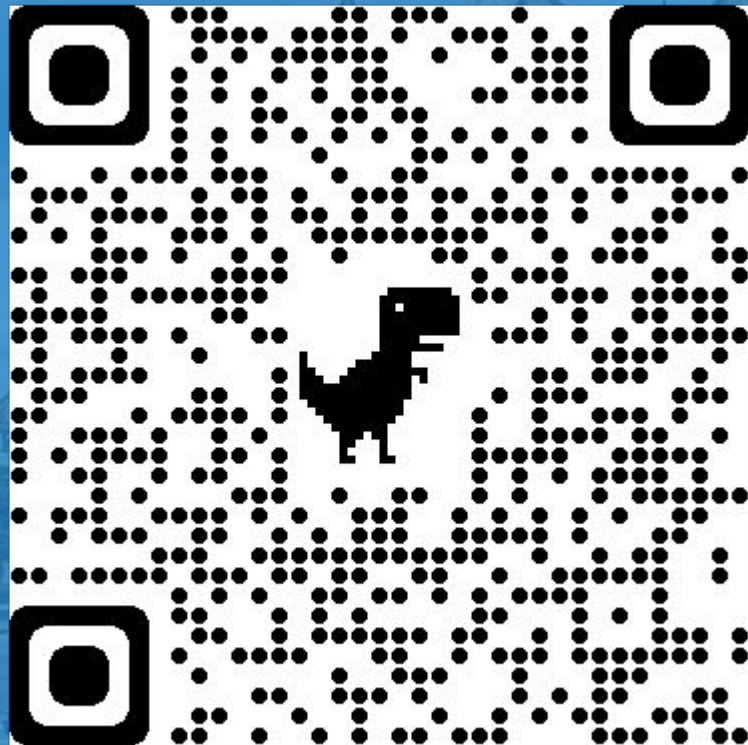
Model	Political Bias		Fake News Spreader	
	UBERT	User2Vec	UBERT	User2Vec
SVM	66.2%	63.0%	53.9%	61.1%
LogReg	64.7%	62.8%	58.6%	59.8%
RnFor	64.9%	63.5%	49.7%	61.3%

- Comparison of different user embeddings for baseline models
- Both political bias and fake news spreaders detection task

- Ablation study over psycho-linguistic features
- Both political bias and fake news spreaders detection task

Model	Political Bias			Fake News Spreader		
	LIWC	BFM	Both	LIWC	BFM	Both
SVM	55.1%	38.8%	61.0%	56.2%	51.0%	53.9%
LogReg	<u>63.6%</u>	51.5%	<u>63.9%</u>	<u>58.3%</u>	55.1%	<u>58.3%</u>
RnFor	56.6%	<u>54.8%</u>	61.7%	55.9%	<u>58.4%</u>	54.8%

- 
- The background of the slide is a photograph of the Philipps-Universität Marburg, showing the main church building and a street in front of it. A semi-transparent blue rectangle is overlaid on the image, containing the list of bullet points.
- We introduced FACTOID a user level factuality and political bias dataset
 - We provide fine-grained scores about the users' factuality and political bias
 - Utilize different user representation techniques to classify them
 - Graph modeling of the users using their social media activity
 - Qualitative analysis of the impact of various psycho-linguistic features



Thank you for
your attention!

SCAN THIS TO ACCESS OUR DATA!!!

