Countering Harmful Online Communication in Brazil: Predicting Fine-Grained Factuality of News and Offensive Context of Social Media Comments

Francielle Vargas

University of São Paulo

November 17, 2023

Leibniz-Institut für Sozialwissenschaften

Harmful Communication in Brazil: **Contextualization**

- During the election period in 2018, denunciations against *xenophobia* had an increase of **2,369.5%**; *public incitement to violence and crimes against life*, **630.52%**; *neo-nazism*, **548.4%**; *homophobia*, **350.2%**; *racism*, **218.2%**; and *religious intolerance*, **145.13%** (Safetnet, 2018).

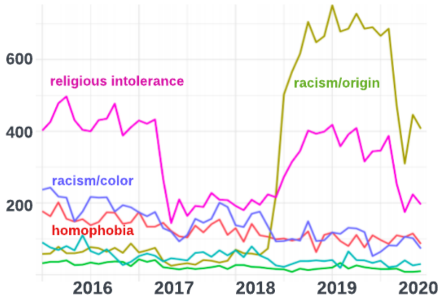


Figure: Hate crimes occurrence in São Paulo from 2016 to the beginning of 2020.

Harmful Communication in Brazil: **Contextualization**

- From 1990 to 2019 there was a **543%** increase in number of protestant churches (BBC Brazil, 2023).
- The Bolsonaro government (2019-2022) was marked by **conservative narratives** (e.g., “**family values**” and “**religious beliefs**“ against “**immorality**“).

Figure: ***“God, country and family”*** was the main slogan used by former Brazilian President Bolsonaro during his electoral campaign and mandate.

Harmful Communication in Brazil: **Cycle**

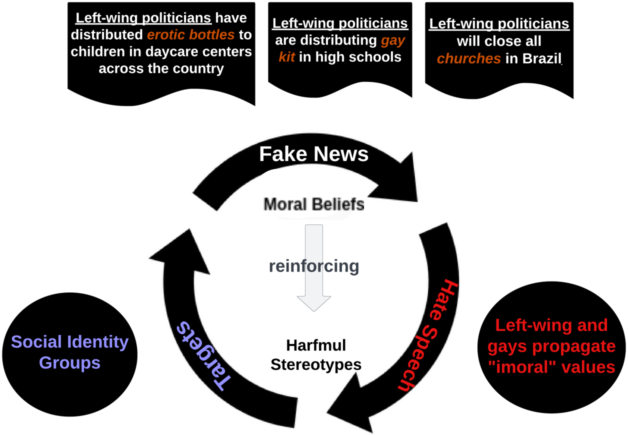


Figure: Harmful communication cycle in Brazil[1].

[1] **Social identity** is a theory of social psychology that offers a motivational explanation for in-group bias.Social identity is a theory of social psychology that offers a motivational explanation for in-group bias.

Harmful Communication in Brazil: **Challenges**

- **Data resources** and **methods** mostly available for the **English** language.
- Towards **addressing the challenges** of the automated fact-checking and hate speech detection.

1. **Hate Speech Detection**:
    - Inaccurate **definition** for offensiveness and hate speech (Davidson et al., 2017).
    - Missing **contextual (cultural)** information (Davidson et al., 2019).
    - Scarce consideration of their **social bias** (Davani et al., 2023)
2. **Automated Fact-Checking and News Credibility Verification**:
    - Fact-checking organizations (e.g. PolitiFact) have provided **lists of unreliable news articles and media sources** (Baly et al., 2018), and most of them address **document-level analysis** of media outlet. Nevertheless, each news article comprises multiple sentences that may contain **factual information**, **bias**, and **fake content**.
    - Automated fact-checking and news credibility verification at scale require **accurate prediction**.

Hate Speech Detection: **Methods and Resources**

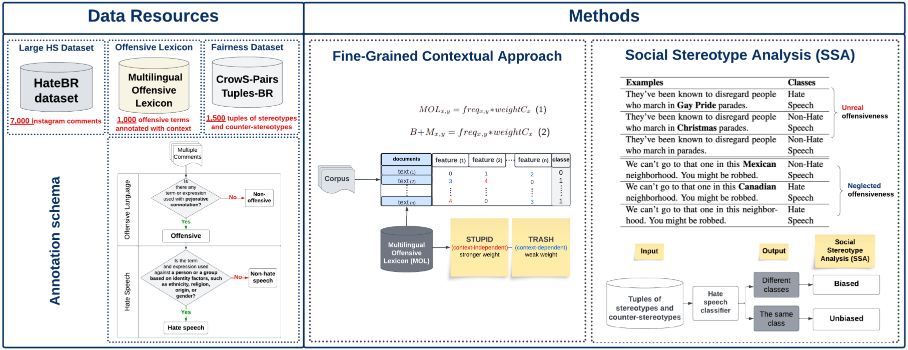


Figure: Data resources and methods for hate speech detection.

Hate Speech Detection: **Results**

Tasks	Features set	Class	Precision				Recall				F1-Score			
			NB	SVM	MLP	LSTM	NB	SVM	MLP	LSTM	NB	SVM	MLP	LSTM
Task 1: Offensive Language Detection	POS+S	0	0.50	0.51	0.47	0.49	0.41	0.39	0.51	0.37	0.45	0.44	0.49	0.42
		1	0.50	0.51	0.54	0.49	0.50	0.64	0.51	0.62	0.59	0.57	0.52	0.55
		Avg	0.50	0.51	0.51	0.49	0.50	0.51	0.51	0.49	0.50	0.50	0.51	0.49
	BOW	0	0.85	0.82	0.92	0.83	0.86	0.96	0.81	0.89	0.86	0.88	0.81	0.86
		1	0.86	0.95	0.79	0.88	0.85	0.79	0.90	0.81	0.85	0.86	0.90	0.85
		Avg	0.85	0.88	0.86	0.85	0.85	0.87	0.86	0.85	0.85	0.87	0.84	0.85
	MOL	0	0.74	0.78	0.94	0.79	0.97	0.96	0.77	0.94	0.84	0.86	0.85	0.86
		1	0.95	0.94	0.72	0.93	0.66	0.73	0.93	0.75	0.78	0.82	0.81	0.83
		Avg	0.85	0.86	0.83	0.86	0.81	0.84	0.85	0.84	0.81	0.84	0.81	0.84
	B+M	0	0.84	0.84	0.91	0.86	0.93	0.94	0.83	0.85	0.88	0.88	0.87	0.85
		1	0.93	0.93	0.81	0.85	0.83	0.81	0.90	0.86	0.88	0.87	0.86	0.85
		Avg	**0.89**	0.88	0.86	0.85	**0.88**	**0.88**	0.87	0.85	**0.88**	0.86	0.86	0.85
Task 2: Hate Speech Detection	POS+S	0	0.52	0.49	0.42	0.52	0.48	0.78	0.53	0.47	0.50	0.60	0.47	0.50
		1	0.52	0.47	0.63	0.52	0.56	0.20	0.52	0.57	0.54	0.28	0.57	0.54
		Avg	0.52	0.48	0.53	0.52	0.52	0.49	0.53	0.52	0.52	0.44	0.52	0.52
	BOW	0	0.62	0.84	0.43	0.85	0.82	0.42	0.82	0.37	0.70	0.55	0.57	0.54
		1	0.73	0.61	0.91	0.61	0.49	0.92	0.61	0.93	0.59	0.73	0.73	0.73
		Avg	0.68	0.72	0.67	0.73	0.66	0.67	0.72	0.66	0.65	0.64	0.65	0.64
	MOL	0	0.61	0.62	0.58	0.60	0.74	0.80	0.68	0.93	0.67	0.69	0.63	0.73
		1	0.67	0.71	0.73	0.84	0.53	0.50	0.63	0.38	0.59	0.59	0.68	0.52
		Avg	0.64	0.66	0.66	0.72	0.64	0.65	0.66	0.65	0.63	0.64	0.66	0.63
	B+M	0	0.79	0.77	0.93	0.71	0.78	0.93	0.79	0.89	0.78	0.84	0.86	0.79
		1	0.78	0.92	0.76	0.85	0.79	0.72	0.92	0.64	0.79	0.80	0.83	0.73
		Avg	0.78	0.84	**0.85**	0.78	0.78	0.83	**0.86**	0.77	0.78	0.82	**0.85**	0.76

Figure: Fine-grained contextual approach for hate speech detection.

Hate Speech Detection: **Results**

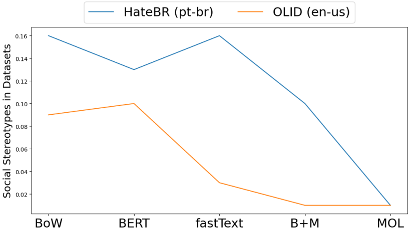


Figure: SSA in different datasets.

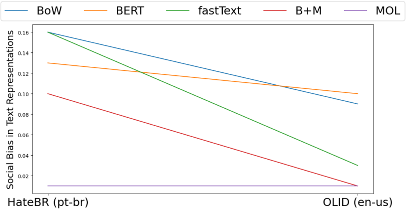
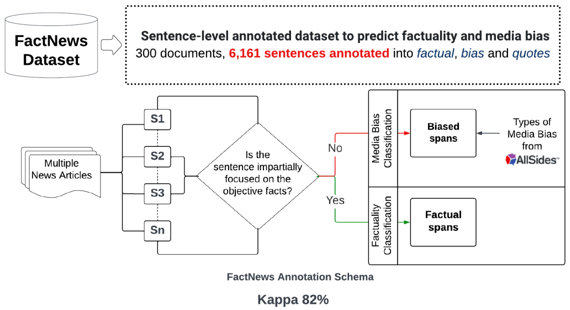


Figure: SSA in ML learning methods.

Automated Fact-Checking: **Methods and Resources**

Data Resource	Method



Sentence-Level Factuality Prediction

N.	Sentence-level news article	Label
Title	President **lowers** Brazil's image with repeated misinformation and does not receive attention from global leaders.	Biased
S1	President Jair Bolsonaro **touch a sore point of** Europeans when he pointed out that the increased use of fossil fuels is a **serious** environmental setback, in his opening speech at the UN General Assembly, Tuesday (20).	Biased
S2	Germany received criticism at the UN for the investment agreement with Senegal for the production of gas in the African country.	Factual
S3	"This constitutes a serious setback for the environment", he said, referring to the Europeans	Quotes
S4	However, Bolsonaro signed measures contrary to environmental protection during the four years of the Brazilian government.	Factual
S5	There is **a huge difference** between speaking at the UN and being heard at the UN.	Biased

Figure: A data resource and method for fact-checking.

Media Bias Examples

Figure: Types of Media Bias Defined by AllSides[2].

[2] https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias

Automated Fact-Checking: **Results**

Description		Folha de São Paulo			Estadão			O Globo			All
		factual	**quotes**	**biased**	**factual**	**quotes**	**biased**	**factual**	**quotes**	**biased**	
#Articles		100			100			100			300
#Sentences		1,494	450	231	1,428	483	182	1,320	458	145	6191
#Words		30,374	7,946	5,177	30,589	8,504	4,002	25,505	7,740	3,195	123,032
Avg Sentences/Article		14.94	7.03	3.78	14.28	7.00	3.19	13.20	7.15	2.84	8.15
Avg Words/Sentences		**20.33**	17.65	**22,41**	**21,45**	17,60	**21,98**	**19,32**	16,89	**22,03**	19,96
Body/Title	*Body*	1,337	440	207	1,218	473	162	1,089	441	131	5,498
	Title	157	10	24	210	10	20	231	17	14	693
Domains	*Political*	912	340	130	870	352	106	748	351	64	3,873
	World	224	48	31	224	49	27	216	32	29	880
	Sports	100	23	34	124	25	29	98	18	39	490
	Daily	132	11	2	98	7	4	148	7	4	413
	Culture	98	26	32	72	42	15	77	45	5	412
	Science	28	2	2	40	8	1	33	5	4	123
Part-of-speech (Avg)	*Noun*	4.85	4.09	**5.72**	5.21	4.12	**5.60**	4.59	3.82	**5.19**	4.79
	Verb	2.20	2.55	**2.60**	2.28	2.51	**2.53**	2.00	2.44	**2.57**	4.18
	Adjective	1.03	1.03	**1.32**	1.11	1.08	**1.32**	0.94	0.97	**1.48**	1.14
	Adverb	0.67	0.82	**0.93**	0.67	0.94	**0.90**	0.59	0.90	**0.94**	0.81
	Pronoun	0.52	1.02	**0.73**	0.51	0.97	**0.56**	0.47	0.90	**0.59**	0.69
	Conjunction	0.51	0.55	**0.61**	0.54	0.57	**0.73**	0.51	0.88	**0.70**	0.62
Emotions (Avg)	*Happiness*	0.12	0.22	**0.20**	0.16	0.28	**0.26**	0.13	0.28	**0.22**	0.20
	Disgust	0.03	0.06	**0.05**	0.04	0.06	**0.03**	0.04	0.04	**0.04**	0.04
	Fear	**4.18**	**3.80**	**4.63**	**4.41**	**3.77**	**4.56**	**4.05**	**3.60**	**4.50**	4.16
	Anger	0.05	0.06	**0.13**	0.07	0.07	**0.12**	0.06	0.08	**0.20**	0.09
	Surprise	0.01	0.03	**0.03**	0.01	0.03	**0.05**	0.01	0.02	**0.01**	0.02
	Sadness	**5.86**	**5.71**	**6.52**	**6.17**	**5.55**	**6.48**	**5.56**	**5.40**	**6.19**	5.93
Polarity (Avg)	*Positive*	2.41	3.25	2.93	2.55	3.22	2.95	2.26	3.26	2.96	2.86
	Negative	0.05	0.06	0.05	0.07	0.10	0.09	0.06	0.07	0.06	0.06
	Neutral	9.55	9.77	10.93	9.92	9.52	11.03	8.91	9.28	10.56	9.94

Table: FactNews dataset statistics.

Automated Fact-Checking: **Results**

- The distribution of **factuality** is *constant* across different domains.
- The distribution of **bias** varies according to the domain and media outlet.

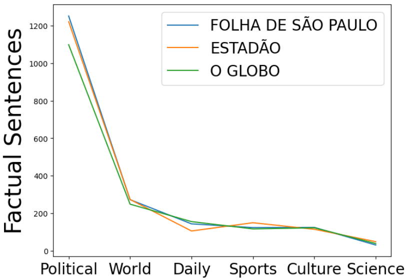


Figure: The cross-domain distribution of factual and biased sentences.

Automated Fact-Checking: **Results**

Sentence-Level Factuality	Precision	Recall	F1-Score
BERT fine-tuning	0.89	0.89	**0.88**
Part-of-speech	0.77	0.77	0.76
TF-IDF	0.81	0.69	0.66
Polarity-lexicon	0.63	0.62	0.62
Emotion-lexicon	0.61	0.61	0.61
Sentence-Level Media Bias	**Precision**	**Recall**	**F1-Score**
BERT fine-tuning	0.70	0.68	**0.67**
Part-of-speech	0.67	0.66	0.66
Polarity-lexicon	0.50	0.50	0.50
Emotion-lexicon	0.53	0.52	0.50
TF-IDF	0.78	0.58	0.48

Figure: Factually Prediction: Evaluation.

Sentence-Level Media Bias Prediction				
Datasets	**Lang**	**Docum.**	**Sent.**	**F1-Score**
BASIL (baseline)	En	300 news	7,984	**0.47**
Biased-sents	En	46 news	966	-
BABE	En	100 news	3,700	0.80
FactNews	Pt	300 news	6,191	**0.67**
Sentence-Level Factuality Prediction				
FactNews (baseline)	Pt	300 news	6,191	**0.88**
Article-Level Factuality Prediction				
MBFC (baseline)	En	1,066 medias	-	**0.58**
MBFC corpus	En	489 medias	-	0.76*

Figure: Result analysis.

Fact-checking and Hate Speech Detection Systems

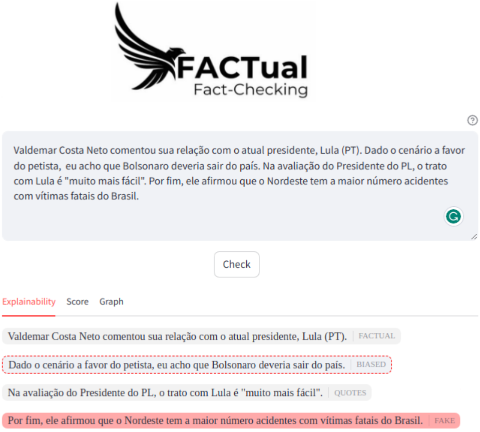


Figure: Automated Fact-Checking.

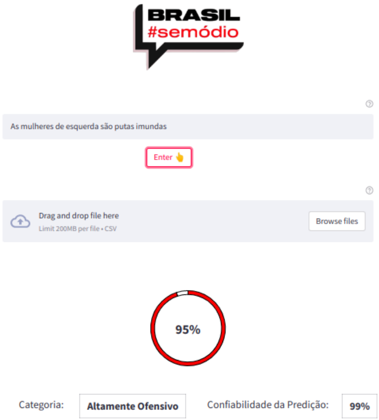
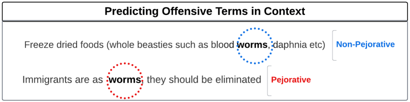


Figure: Automated Offensiveness Analysis.

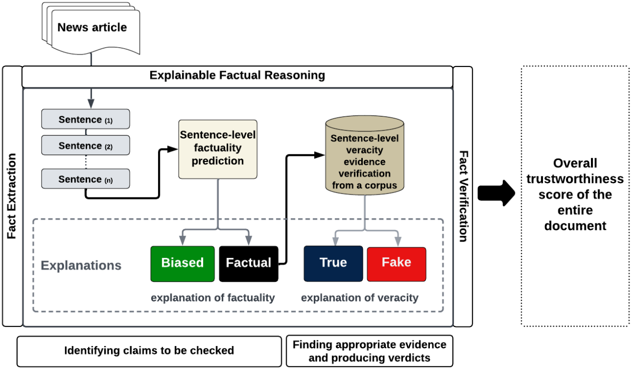
Ongoing Research

1. **Hate Speech**:



HausaHate
A Hate Speech Dataset of Hausa Facebook Comments

2. **Automated Fact-Checking**:



Thank you!

francielleavargas@usp.br

Take a picture to access the papers, datasets, models, and systems

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.