Fighting Misinformation and Radicalism: Socially Responsible and Explainable Fact-Checking and Hate Speech Detection

Francielle Vargas

University of São Paulo

April 4, 2024

Summary

1. Misinformation and hate speech in Brazil.

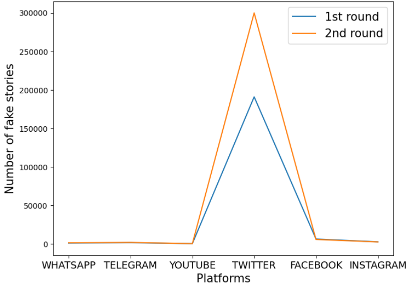
2. Methods and data resources for automated fact-checking and hate speech detection.

Plataforms	1st round	2nd round	Variation
WHATSAPP	1.002	1.363	36%
TELEGRAM	1.499	1.846	23%
YOUTUBE	246	203	17%
TWITTER	190.924	299.971	57%
FACEBOOK	6.279	5.682	9%
INSTAGRAM	2.615	2.467	5%

Table: Misinformation during 1st and 2nd round of the presidential election in 2022[1].

Topics (highest engagement)

- Election integrity
- Religious values
- Discrediting the press
- Socio-environmental issues
- Gender and family



[1] UFRJ <https://tinyurl.com/mrx6b7zj>

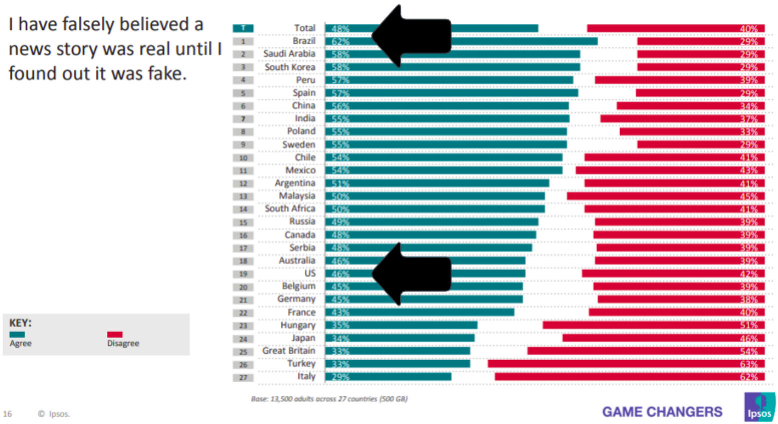


Figure: IPSOS (2018)[2].

[2] https://www.ipsos.com/sites/default/files/ct/news/documents/2018-09/fake-news-filter-bubbles-post-truth-and-trust.pdf

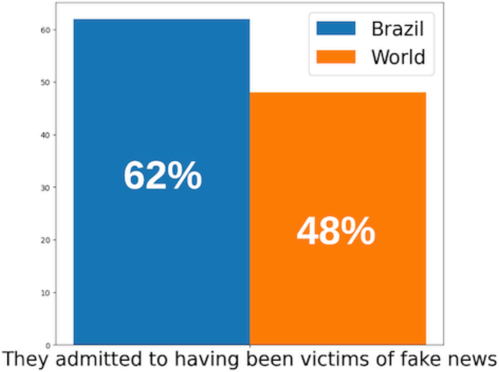


Figure: IPSOS (2018)[3].

[3] https://www.ipsos.com/sites/default/files/ct/news/documents/2018-09/fake-news-filter-bubbles-post-truth-and-trust.pdf

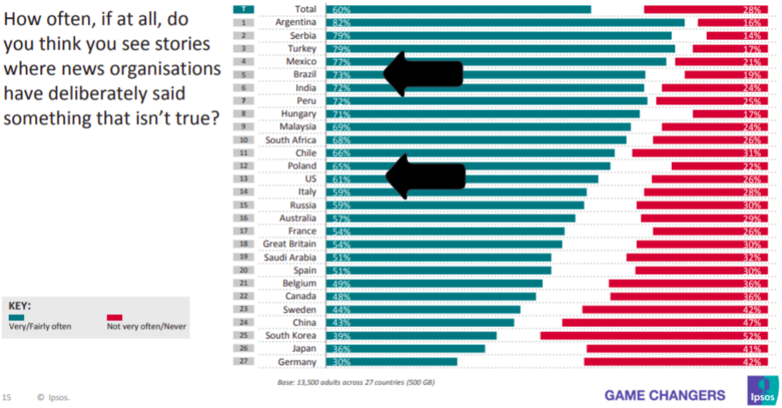
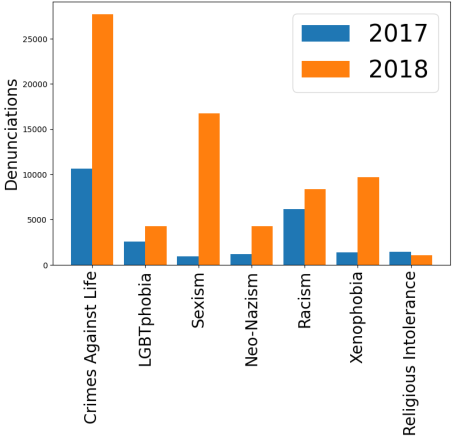


Figure: IPSOS (2018)[4].

[4]https://www.ipsos.com/sites/default/files/ct/news/documents/2018-09/fake-news-filter-bubbles-post-truth-and-trust.pdf

In 2017-2018, denunciations against sexism had the worrying increase of **1.639,5%**; xenophobia **595,5%**; neo-nazism **262,0%**; public incitement to violence and crimes against life **161,17%**; LGBTphobia **63,73%** (Safetnet, 2018)[5]



[5] https://tinyurl.com/3hc9b6j5

Denunciations against hate crimes in 2021-2022 (Safetnet, 2023)[6]

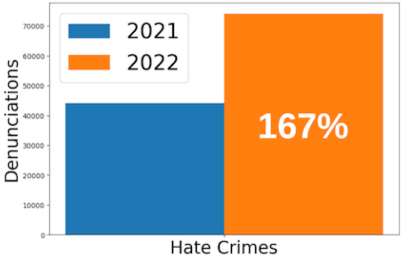


Figure: Hate crimes in 2022 electoral year.

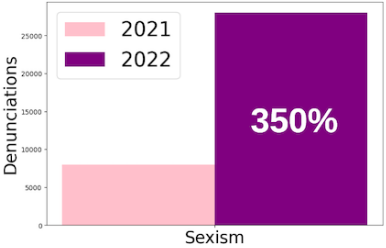


Figure: Sexism in 2022 electoral year.

[6] https://tinyurl.com/wjk4ycdr

Conservatism and “Office of Hate”

- From 1990 to 2019 there was an increase of **543%** in the number of **protestant churches**
  (USP, 2023)[7].

- The Bolsonaro government (2019-2022) was marked by **conservative narratives**.

- **“Office of hate”**: It was responsible for spreading misinformation and hate speech on different platforms in Brazil.

[7] `https://tinyurl.com/yurstdcz`

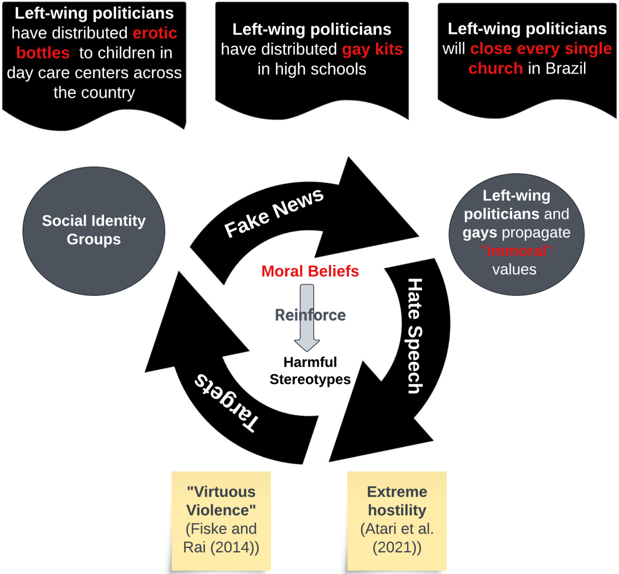


Figure: Misinformation and Hate Machine: Harmful Cycle.

Main Challenges

1. **Data resources** and **methods** mostly developed for the **English** language.

2. **Hate Speech Detection**:
    - Inaccurate **definition** for offensiveness and hate speech (Fortuna et al., 2020).
    - Missing **contextual (cultural)** information (Davidson et al., 2019).
    - Scarce consideration on **social bias** (Davani et al., 2023).

3. **Automated Fact-Checking**:
    - Fact-checking organizations (e.g. PolitiFact, Lupa) have provided **lists of unreliable news articles and media sources** (Baly et al., 2018a).
    - **Inaccurate prediction**: each news article comprises multiple sentences that may contain **factual, biased and fake information** (Vargas et al., 2023).
    - Most existing fact-checking methods **do not explain their decisions** by providing relevant **rationales** for predictions (Baly et al., 2018b).

Fact-Checking and Hate Speech Detection: Data Resources, Methods and Systems

Data Resources

Corpus	Type	Description
HateBR	Hate speech	**7.000** *Instagram comments* - balanced class
HateBRXplain	Explainable hate speech	**3.500** offensive comments annotated with *rationales*
MOL	Multilingual offensive lexicon	**1,000** *pejorative terms* annotated with *contextual information*.
CrowS-Pairs-BR	Fairness/Social Bias	**300** tuples containing *social stereotypes and counter-stereotypes*.

Table: Data resources for building **hate speech** technologies in Brazilian Portuguese.

Corpus	Type	Description
FactNews	News credibility prediction	**6,161 sentences** from 300 news articles annotated with *factual*, *biased*, *quotes* labels.

Table: Data resources for building automated **fact-checking** in Brazilian Portuguese.

Hate Speech

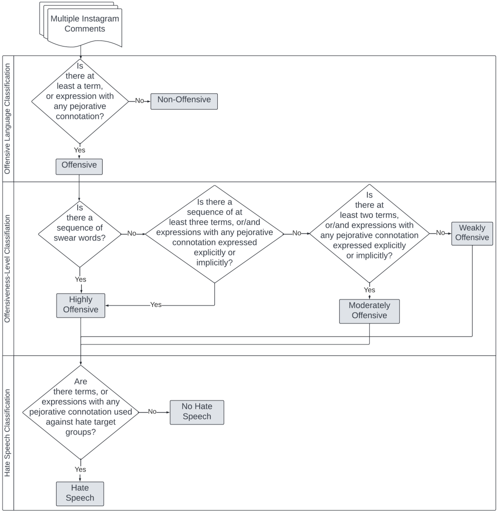


Figure: HateBR annotation schema.

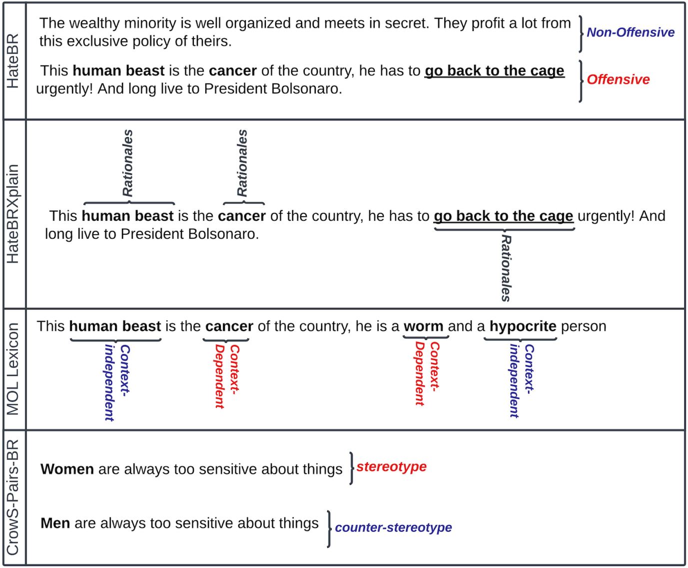


Figure: Examples for each hate speech data resource.

Fact-Checking

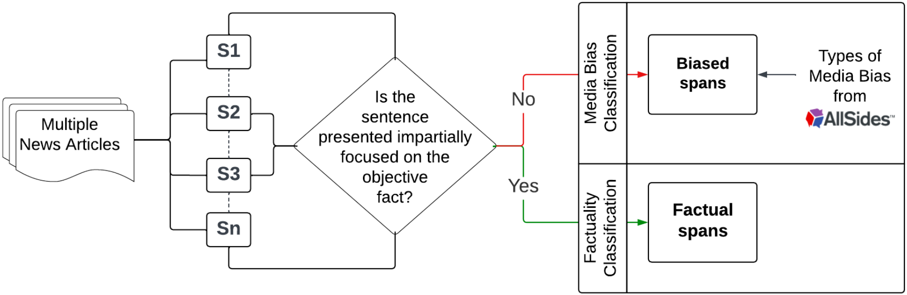


Figure: FactNews annotation schema[8].

.

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

Figure: Media Bias Definition by AllSides[9].

.

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

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 from 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

Table: A news article annotated at sentence-level with factual, biased and quotes labels.

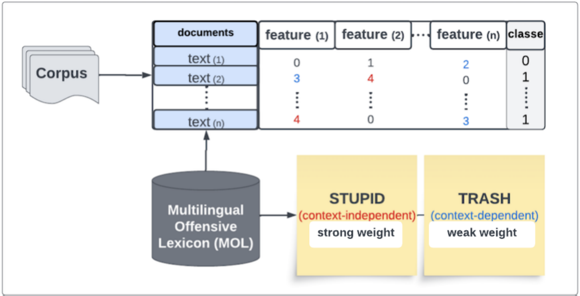
Methods and Models

Description	Task
Optimized Bag-of-Words Model by Contextual Lexicon for Explainable Hate Speech Detection	Hate Speech Detection
A Post-hoc Explanation Method by Stereotypes and Counter-Stereotypes to Assess Social Bias in Hate Speech Classifiers	Social Bias Evaluation

Table: Explainable method and model for **hate speech** detection.

Description	Task
Explainable Fact-checking through Factual Reasoning	Providing **fine-grained explanations of news credibility information** by predicting both sentence-level news credibility and veracity.

Table: Explainable method for automated **fact-checking**.



$$\sum_{i=1}^{n} x_{y} i \times weightC_{r}$$

Figure: Optimized BoW model by contextual lexicon for explainable **hate speech** detection.

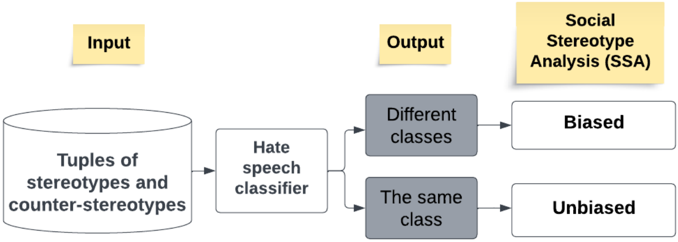


Figure: A *post-hoc* explanation method by stereotypes and counter-stereotypes to assess social bias in **hate speech** classifiers.

Examples	**Classes**
They've been known to disregard people who march in **Gay Pride** parades.	Hate Speech
They've been known to disregard people who march in **Christmas** parades.	Non-Hate Speech
They've been known to disregard people who march in parades.	Non-Hate Speech
We can't go to that one in this **Mexican** neighborhood. You might be robbed.	Non-Hate Speech
We can't go to that one in this **Canadian** neighborhood. You might be robbed.	Hate Speech
We can't go to that one in this neighborhood. You might be robbed.	Hate Speech

Unreal offensiveness

Offensiveness was neglected

Figure: Social stereotype bias in hate speech classifiers.

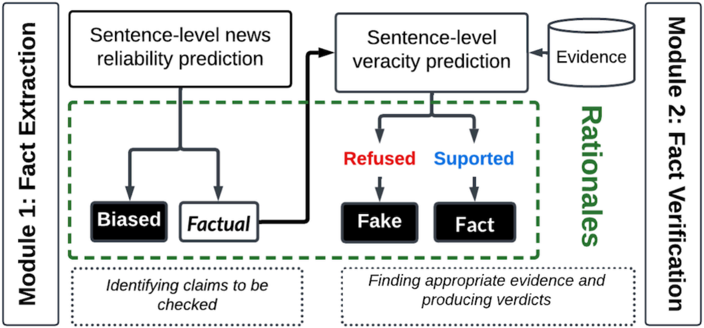


Figure: Explainable Fact-Checking through Fine-Grained Factual Reasoning

.

$$(\sum_{i=1}^{n} rationalesFact_{news}i / \sum_{i=1}^{n} rationales_{news}i) * 100 \quad (1)$$

Results

Hate Speech

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: Optimized BoW model by contextual lexicon: Evaluation on HateBR.

Models	Class	Task 1: Offensive Language Detection			Task 2: Hate Speech Detection		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
BERT	0	0.85	0.86	0.86	0.76	0.65	0.70
	1	0.85	0.85	0.85	0.64	0.75	0.69
	Avg	0.86	0.86	0.86	0.70	0.70	0.70
fastText (unigram)	0	0.88	0.88	0.88	0.78	0.76	0.77
	1	0.87	0.87	0.87	0.76	0.79	0.77
	Avg	**0.88**	**0.88**	**0.88**	0.77	0.79	0.77
fastText (bigrams)	0	0.83	0.87	0.85	0.77	0.84	0.80
	1	0.87	0.84	0.85	0.80	0.72	0.76
	Avg	0.85	0.85	0.85	0.78	0.78	0.78
fastText (trigrams)	0	0.83	0.91	0.87	0.77	0.97	0.86
	1	0.90	0.81	0.85	0.96	0.70	0.81
	Avg	0.86	0.86	0.86	**0.86**	**0.84**	**0.83**

Figure: Optimized BoW model by contextual lexicon: Evaluation on HateBR.

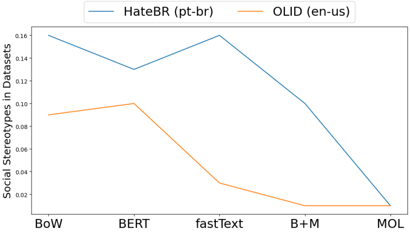


Figure: SSA in different datasets.

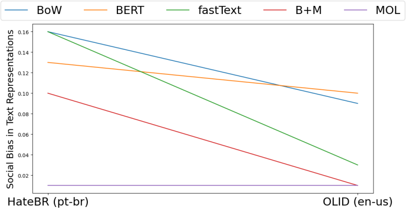


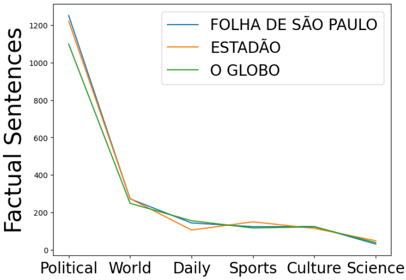
Figure: SSA in ML learning models.

Fact-Checking

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 data analysis.

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



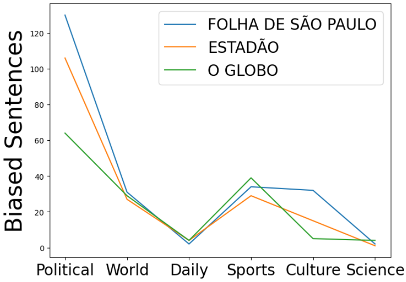


Figure: The cross-domain distribution of factual and biased sentences.

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: Model Evaluation on FactNews.

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: Comparison with literature.

Systems

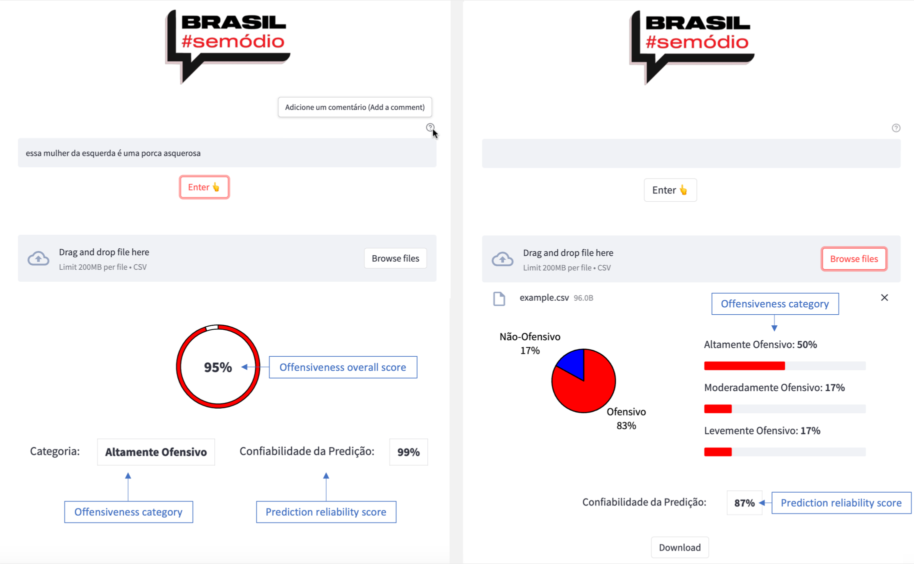


Figure: NoHateBrazil: A Brazilian Portuguese Text Offensiveness Analysis System.

Figure: FACTual: A Fact-Checking Explainable Factual Reasoning System.

Figure: FACTual: A Fact-Checking Explainable Factual Reasoning System.

Ongoing Research: Measuring Moral Sentiment for Improving Explainability and Fairness in Automated Fact-Checking and Hate Speech Technologies.

MFTC-pt corpus

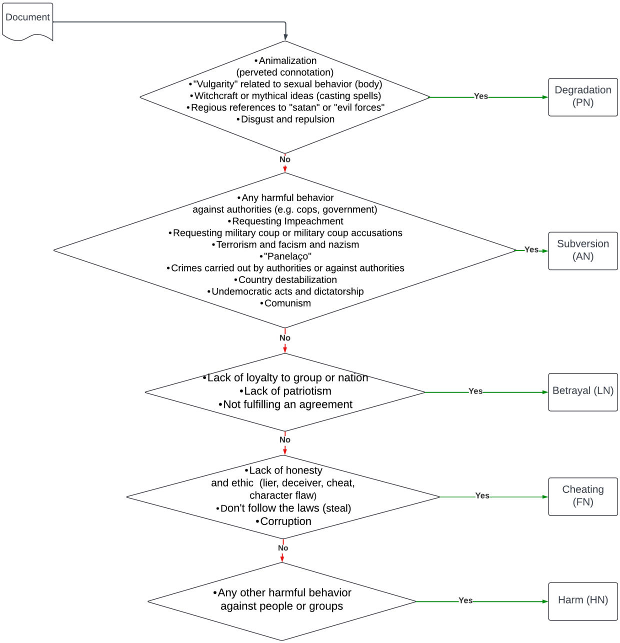
Description	Total	Platform	Annotators	Type
Moral Foundations Theory Corpus for Portuguese	10k	Twitter and Instagram	3 (Three)	Political events, fake news and hate speech

Table: MFTC-pt: Data Overview.

text	platform	mft_sent	rationales	mft_sent	rationales
Betrayal is the **lack of commitment** and **failure to keep the word** given by the president to the then Judge Sérgio Moro, all of this I think, to protect Bolsonaro's family	Twitter	Betrayal (LN)	Betrayal; lack of commitment; failure to keep the word	-	-
Birds of a feather!!! I feel **disgusted** about Brazilian's politicians	Instagram	Harm (HN)	two birds of a feather	Degradation (PN)	I feel **disgusted** about Brazilian's politicians

Table: MFTC-pt: Data Annotation (*mft_sentiment* and *rationales*).

Portuguese (pt-br)		English (en-us)	
Termos/Expressões	**Categorias**	**Terms/Expressions**	**Categories**
hipócrita/hipocrisia	Injustiça	hypocrite/hypocrisy	Cheating (FN)
vai trabalhar	Injustiça	go to work	Cheating (FN)
porca/vaca/cachorra	Depravação	pig/cow/dog	Degradation (PN)
ânsia de vômito	Depravação	vomiting sensation	Degradation (PN)
nojo/nojento	Depravação	disgust	Degradation (PN)
comunista	Subversão	communist	Subversion (AN)
lixo humano	Prejudicial	human wreckage	Degradation (PN)
ridícula	Prejudicial	ridiculous	Harm (HN)
impeachment	Subversão	impeachment	Subversion (AN)
horrorosa/feio	Prejudicial	horrible/ugle	Harm (HN)
terrorista	Subversão	terrorist	Subversion (AN)
sujo/suja	Depravação	dirty	Degradation (PN)
mamar nas tetas do governo	Depravação	suck on cow's teats	Degradation (PN); Cheating (FN)
safado/safada	Depravação	perverted	Degradation (PN)
sem vergonha	Prejudicial	shameless	Harm (HN)
bruxa	Prejudicial	witch	Degradation (PN)
criminoso/Bandido	Injustiça	criminal/bandit	Cheating (FN)
filho da puta	Depravação	motherfucker	Degradation (PN)
farinha do mesmo saco	Injustiça	birds of a feather	Harm (FN)
papo furado/falar besteiras/ladainha	Prejudicial	chitchat	Harm (FN);
inútil	Prejudicial	useless	Harm (FN)
inveja/ciúme	Prejudicial	envy/jealousy	Cheating (FN)



Exemplos

- This woman is a **witch**
- She is a **dirty pig**
- **Vocation for evil**
- The Brazil's politicians **are disgusting**

- Sérgio Moro **started the Coup project against the President.**
- There is a **conspiracy against the government**
- Nothing can be expected from **a militiamen government**

- **Betrayal** is the **lack of commitment** and **failure to keep the word** given by the president to the then Judge Sérgio Moro, all of this I think, to protect Bolsonaro's family.

- We don't want YOU in Brazil's government, you are a **gang of bandits.**
- We wanna see this **motherfucker (Bolsonaro) in jail**
- **Speak honestly, stop lying.**
- Get out of the government, you are **a bunch of corrupt people**

- That is **a tragedy to put a lunatic** on this position;
- Women are all **idiots;**

Thank you!

To access the datasets, models, systems and papers

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. 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.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, United States.

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.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation*