Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection

Francielle Vargas

Dr. Thiago Pardo (ICMC-USP) & Dr. Fabrício Benevenuto (DCC-UFMG)

University of São Paulo

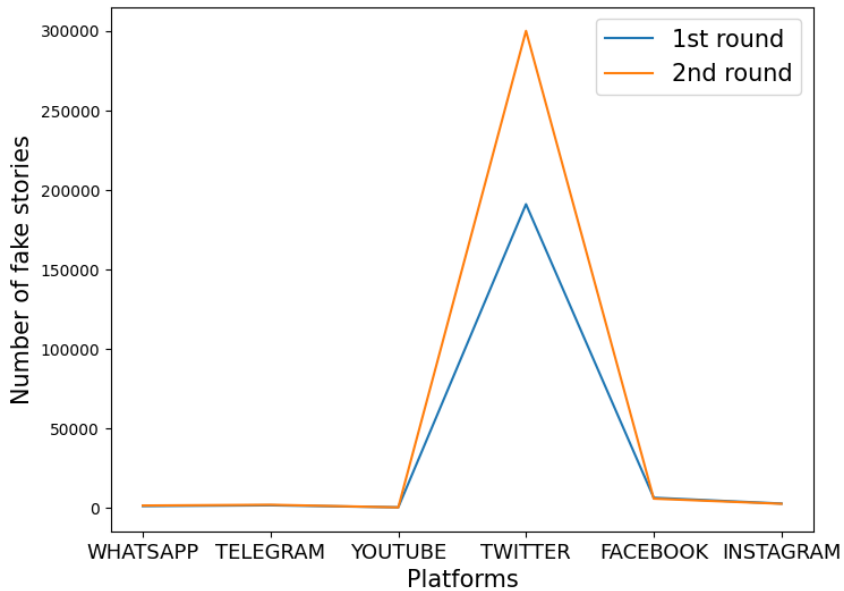
Misinformation and hate speech have a negative impact on society, particularly in conflict-affected areas and politically polarized countries. These issues are fueled by longstanding and ingrained social, cultural, political, ethnic, religious, and other divisions and rivalries, often exacerbated by misinformation through sophisticated belief systems, including propaganda and conspiracy theories (Wardle, 2024).

Platforms	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¿

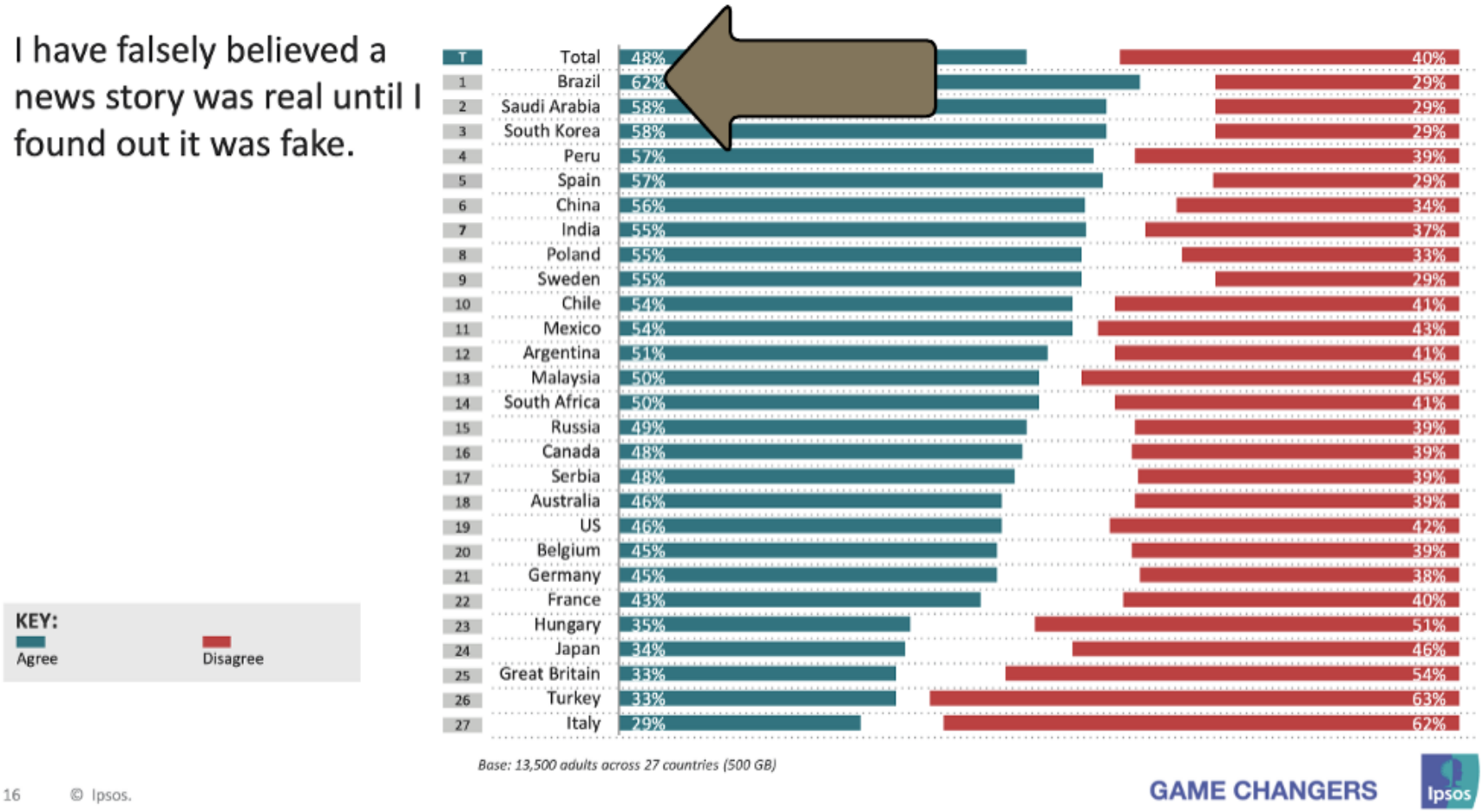
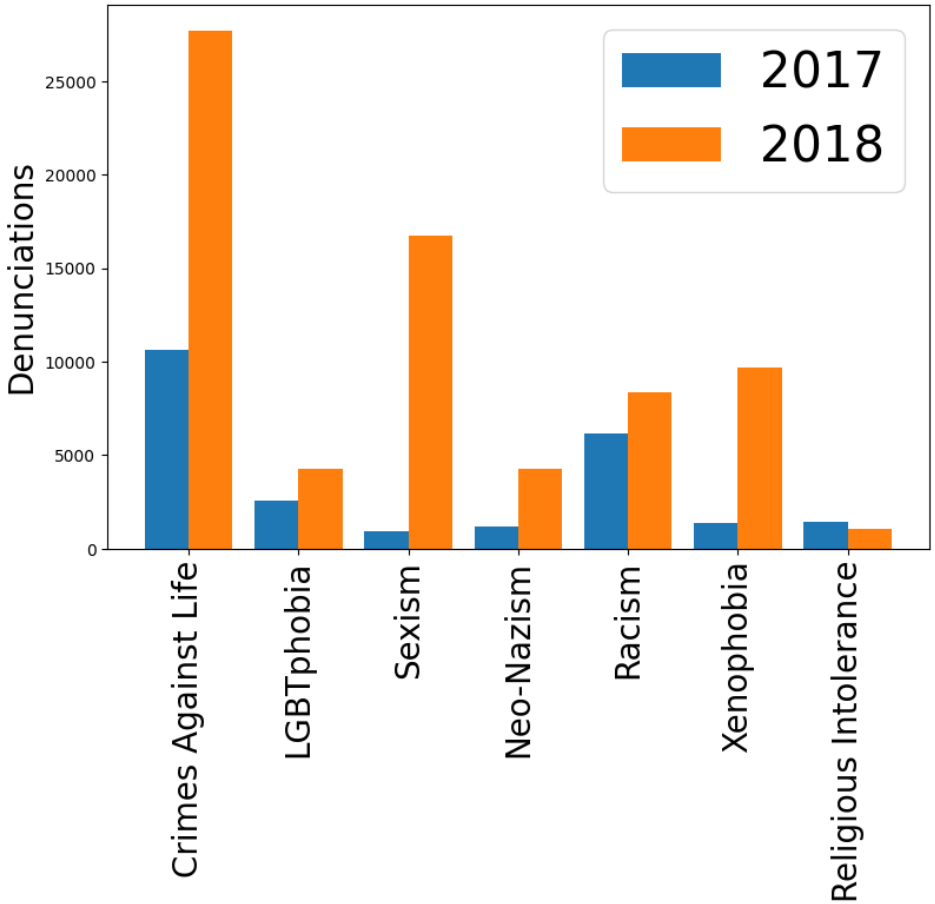


Figure: IPSOS (2018)[2].

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

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%** (Safernet, 2018)[3]



[3] https://new.safernet.org.br/

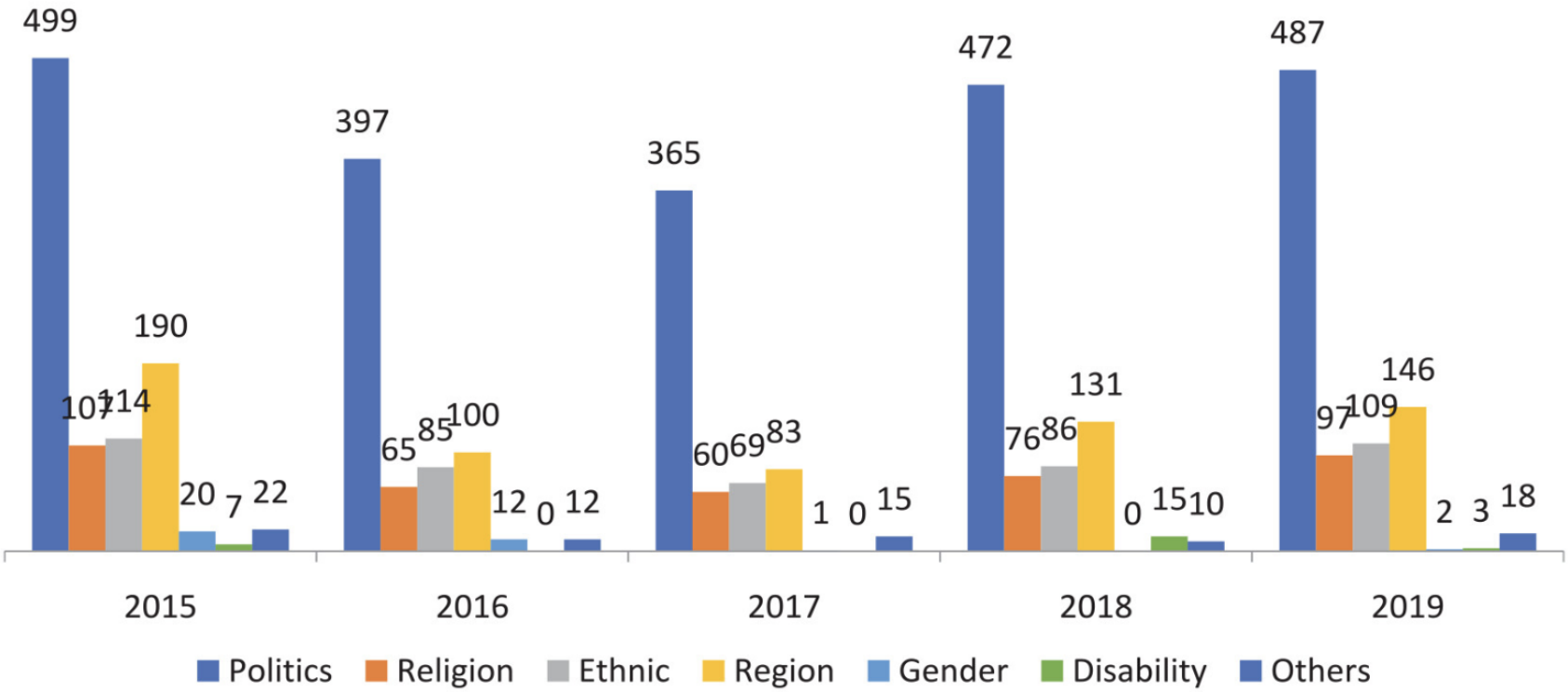


Figure: Most Dominant Themes in Hateful Tweets in Nigeria [4]

.

[4] Ridwanullah, A. O., Sule, S. Y., Usman, B., and Abdulsalam, L. U. (2024). *Politicization of Hate and Weaponization of Twitter/X in a Polarized Digital Space in Nigeria*. Journal of Asian and African Studies, 1(1).

Research Gaps

1. **Datasets** and **resources** 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).
   - **Lack of Transparency**: Scarce consideration on **social bias** (Davani et al., 2023).

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

Research Question

Can research on Explainable AI and Fairness enhance the transparency of black-box models and reduce the risk of negative social impacts?

Contributions

1. **5 (Five) Benchmark Datasets for Low-Resource Languages**:

   HateBR, HateBRXplain, HausaHate, MOL, CrowS-Pairs-BR and FactNews.

2. **4 (Four) Computational Methods**:
   - **B+M**: Contextualized Bag-of-Words with Feature Saliency for Explainable Hate Speech Detection.
   - **SRA**: Supervised Rational Attention for Self-Explaining Hate Speech Detection.
   - **SSA**: A Counterfactual Explanation Approach to Assess Social Bias in Hate Speech Classifiers.
   - **SELFAR**: Sentence-Level Factual Reasoning for Explainable Fact-Checking.

3. **1 (One) Web System**:
   - **NoHateBrazil**: A System for Text Offensiveness Analysis in Brazilian Portuguese.

Benchmark Datasets for Low-Resource Languages

Corpus	Type	Description
HateBR[5]	Hate speech	**7.000** *Instagram comments* - balanced class.
HateBRXplain[6]	Explainable hate speech	**3.500** offensive comments annotated with *human-based rationales*.
HausaHate[7]	Hate speech	**2,000** comments extracted from Western African Facebook pages.
MOL[8]	Context-aware and expert multilingual offensive lexicon	**1,000** *pejorative terms* annotated with *contextual information*.
CrowS-Pairs-BR[9]	Fairness/Social Bias	**300** tuples containing *social stereotypes and counter-stereotypes*.

Table: **Hate speech detection Brazilian Portuguese and Hausa Languages**.

Corpus	Type	Description
FactNews[10]	News credibility prediction	**6,161 sentences** from 300 news articles annotated with *factual*, *biased*, *quotes* labels.

Table: **Automated fact-checking in Brazilian Portuguese**.

[5] LREC 2022
[6] COLING 2025
[7] NAACL 2024 WOAH
[8] Natural Language Processing Journal
[9] RANLP 2023
[10] RANLP 2023

Post-Hoc and Self-Explaining Computational Methods

Contextualized Bag-of-Words with Features Saliency (B+M)

$$MOL_{x,y} = freq_{x,y} * weightC_x \quad (1)$$

$$B+M_{x,y} = freq_{x,y} * weightC_x \quad (2)$$

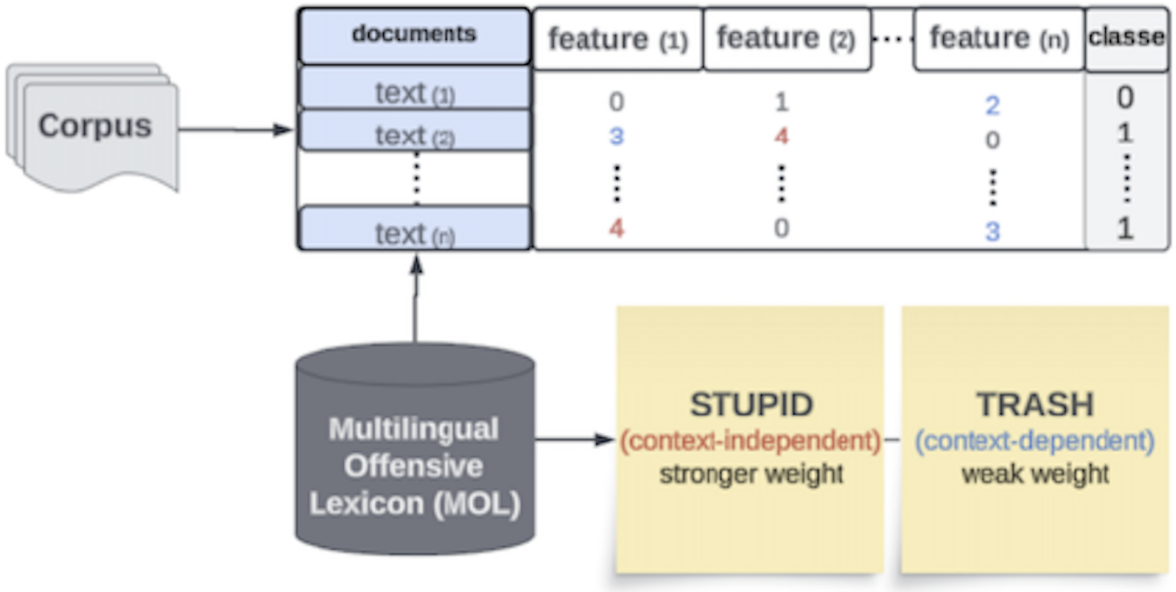


Figure: Francielle Vargas, Isabelle Carvalho, Fabiana R. Góes, Fabrício Benevenuto, Thiago A.S. Pardo. **Contextual-Lexicon Approach for Abusive Language Detection.** *13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1438–1447.*

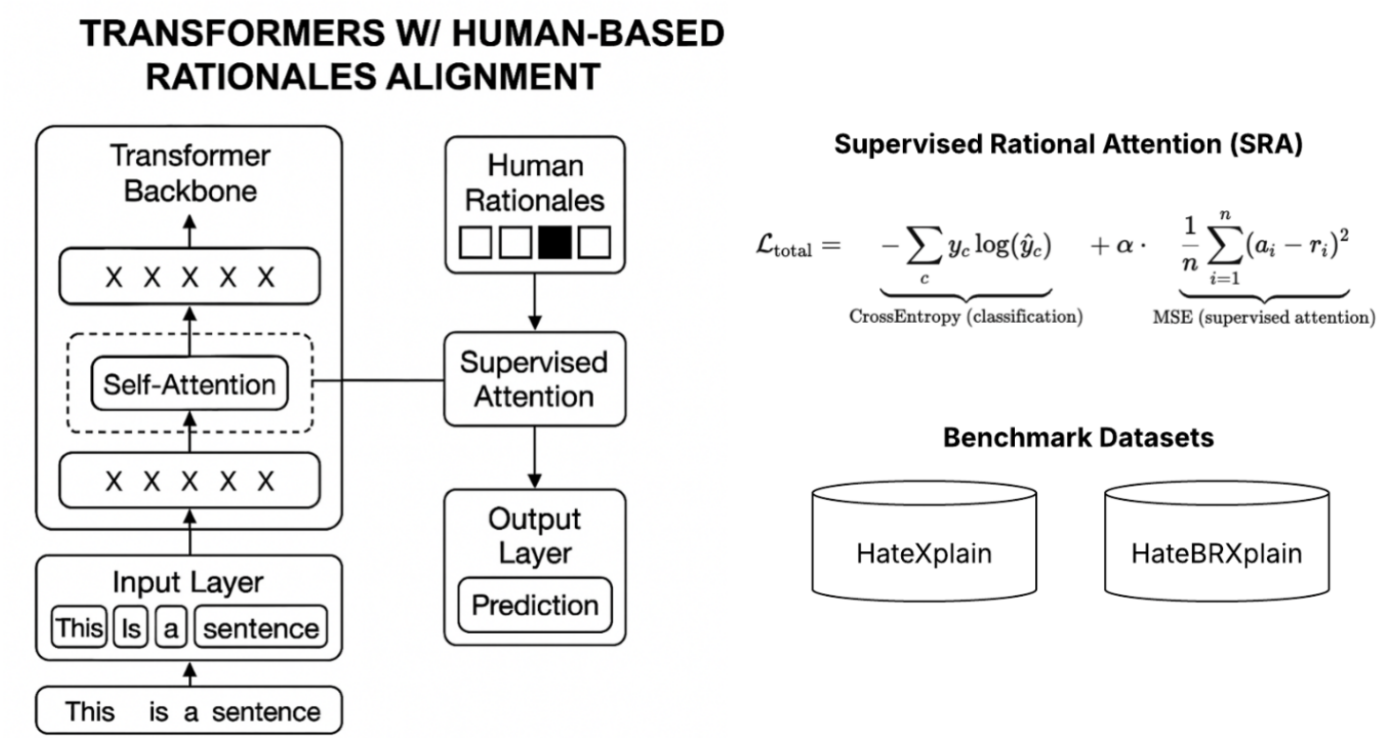


Figure: Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, Ali Ramezani-Kebrya. **Aligning Attention with Human Rationales for Self-Explaining Hate Speech Detection**. *40th Annual AAAI Conference on Artificial Intelligence (AAAI 2026).* (to appear)

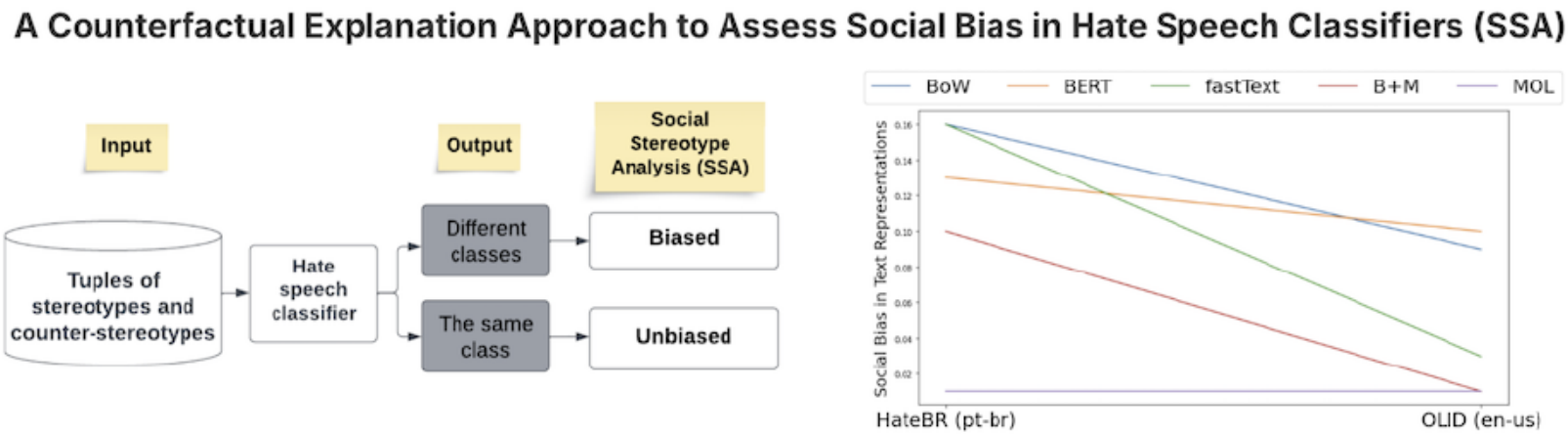


Figure: Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago A.S. Pardo, Fabrício Benevenuto. **Socially Responsible Hate Speech Detection: Can Classifiers Reflect Social Stereotypes?** *14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023). pp. 1187–1196.*

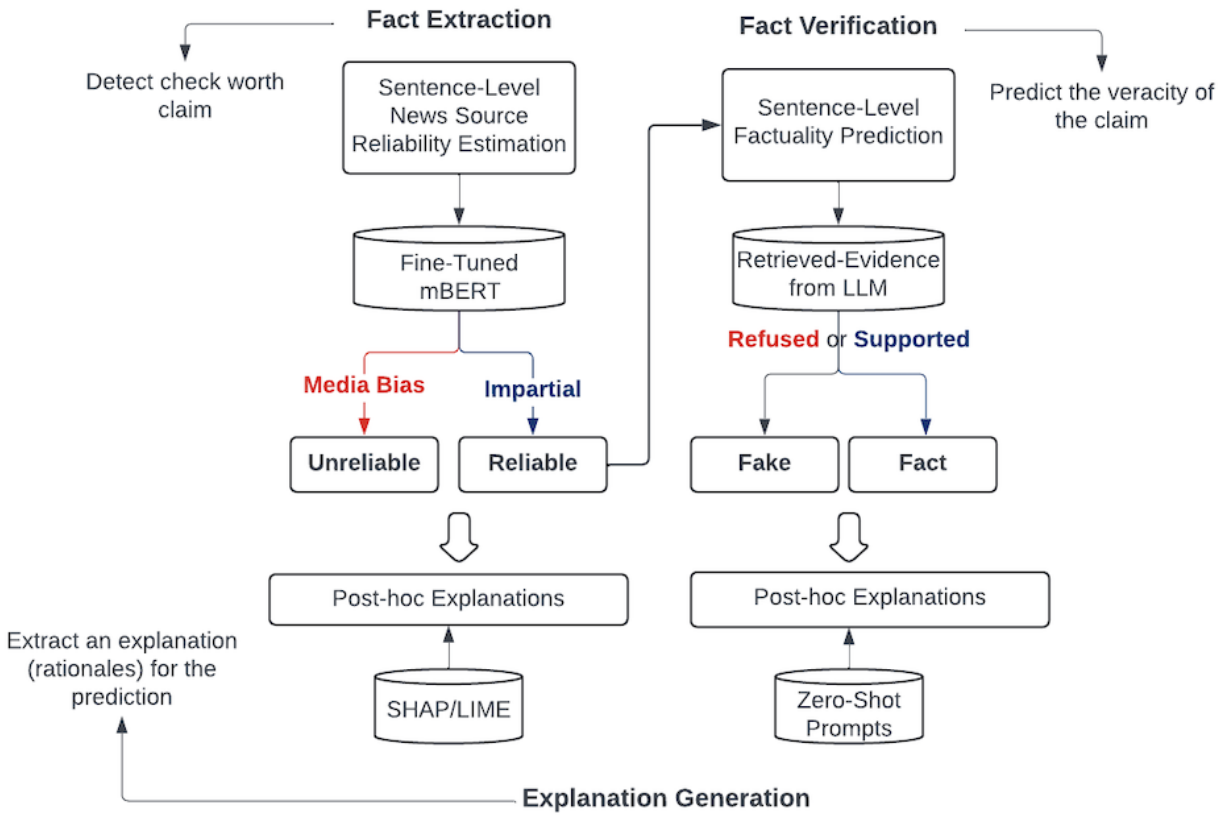


Figure: Francielle Vargas, Kokil Jaidka, Thiago A.S. Pardo, Fabrício Benevenuto. **Predicting Sentence-Level Factuality of News and Bias of Media Outlets.** *14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023).* pp. 1197–1206.

Explainable System

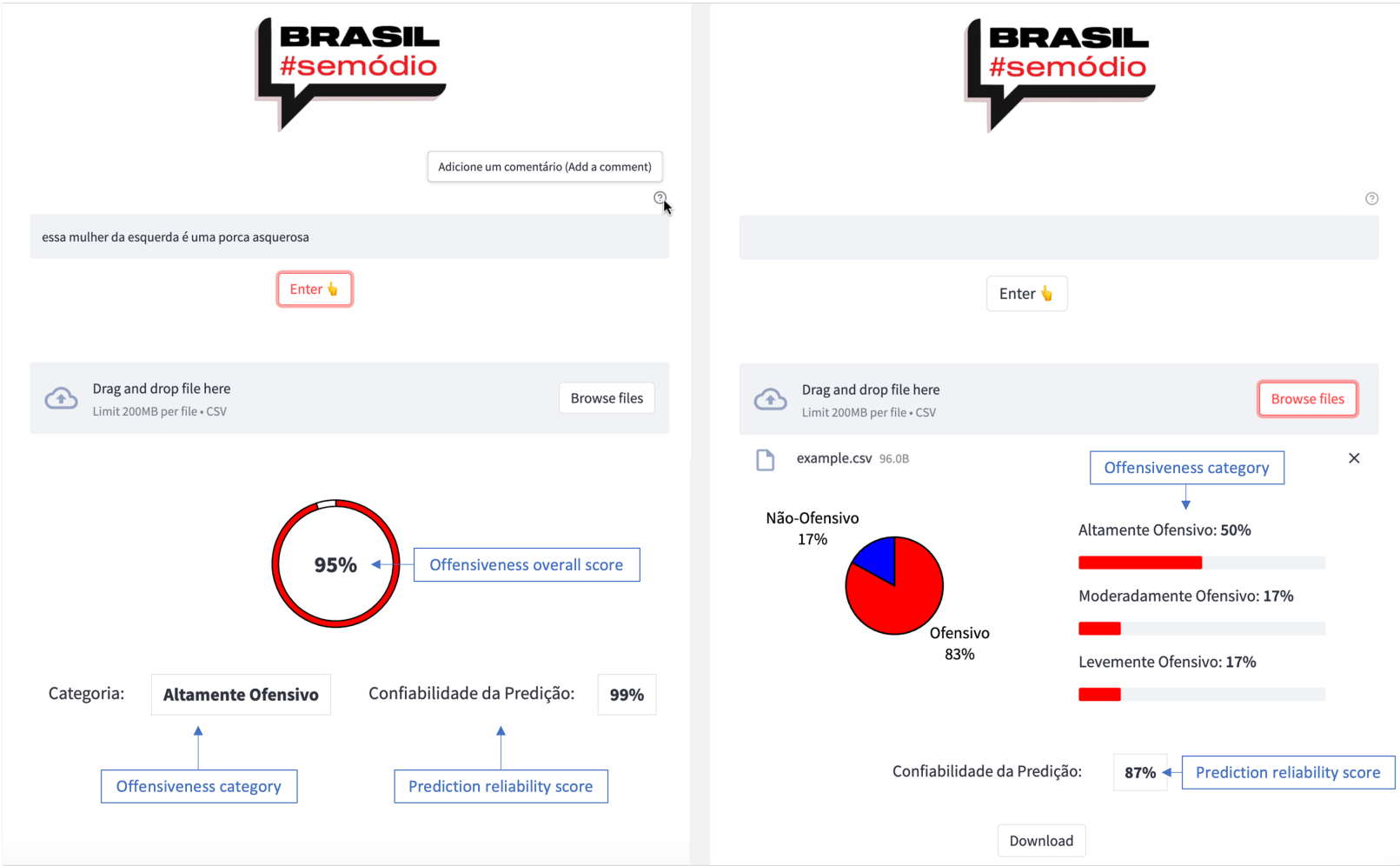


Figure: Francielle Vargas, Isabelle Carvalho, Wolfgang Schmeisser-Nieto, Fabrício Benevenuto, Thiago A.S. Pardo. **NoHateBrazil: A Brazilian Portuguese Text Offensiveness Analysis System**. *14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023).* pp. 1180-1186.

Algorithm 1 Offensiveness Overall Score

procedure GET-OOS($prob$)
  if $mol_{indep} >= 1$ or $mol_{dep} >= 3$ **then**
    $OOS = (90 + score_{prob}) \div 2$
  end if
  if $mol_{dep} == 2$ **then**
    $OOS = (60 + score_{prob}) \div 2$
  end if
  if $mol_{dep} == 1$ **then**
    $OOS = (30 + score_{prob}) \div 2$
  end if
  if $OOS > 0$ and $OOS <= 49$ **then**
    $class = slightly\ offensive$
  end if
  if $OOS >= 50$ and $OOS <= 79$ **then**
    $class = moderately\ offensive$
  end if
  if $OOS >= 80$ and $OOS <= 100$ **then**
    $class = highly\ offensive$
  end if
  return OOS and $class$
end procedure

Figure: The SSO (Offensiveness Overall Score) is a **rules-based algorithm** designed to predict whether a comment is *slightly, moderately, or highly offensive*. Specifically, the SSO uses human-annotated expert rationales from the MOL, embedding contextual knowledge combined with statistical insights to automatically predict the level of text offensiveness. The SSO achieved an accuracy of 0.70.

Results: Performance and Explainability

Models	Precision	Recall	F1
AfriBERTa_base	80.3	80.1	80.2
Afro-XLMR-base	74.8	75.6	74.8
mBERT-cased	74.3	75.1	73.7
XLM-R-base-Hausa	**85.9**	**86.1**	**85.8**

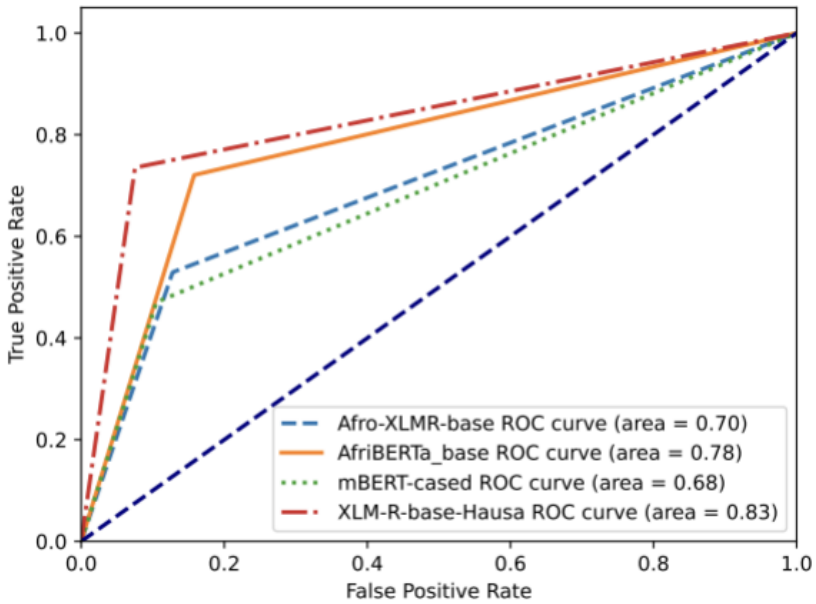


Figure: Evaluating HausaHate: A Baseline for Hate Speech Detection in the Hausa Indigenous Language.

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: Evaluation of **B+M Method** on the HateBR corpus using NB, SVM, MLP and LSTM.

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: Comparison of **B+M Method** with mBERT and fastText.

	FE	FV	SELFAR
class	F1	F1	F1
0	0.85	0.61	0.60
1	0.82	0.81	0.85
Avg	0.85	0.71	**0.72**

Table: Evaluation of **SELFAR**. Note that for FE, the classes are reliable (0) and unreliable (1). In contrast, for FV and SELFAR, the classes are fact/true (0) and fake (1).

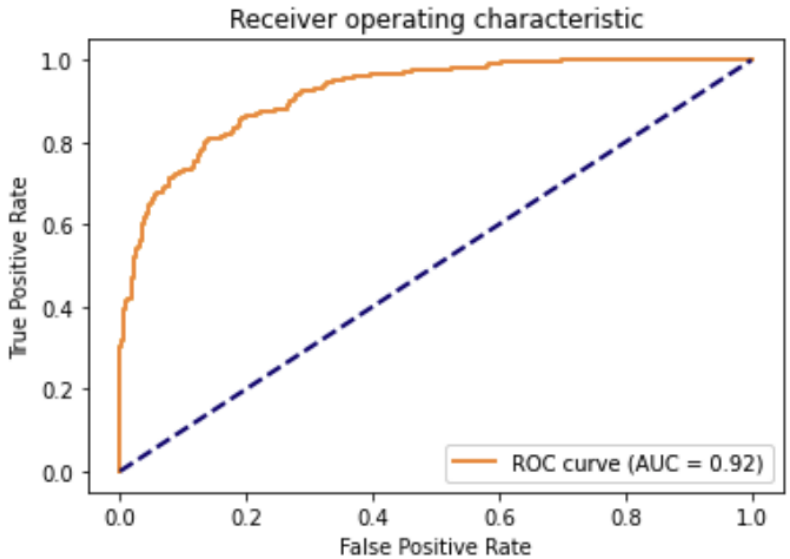


Figure: ROC SELFAR.

Explainability Evaluation

Counterfactual Explanations

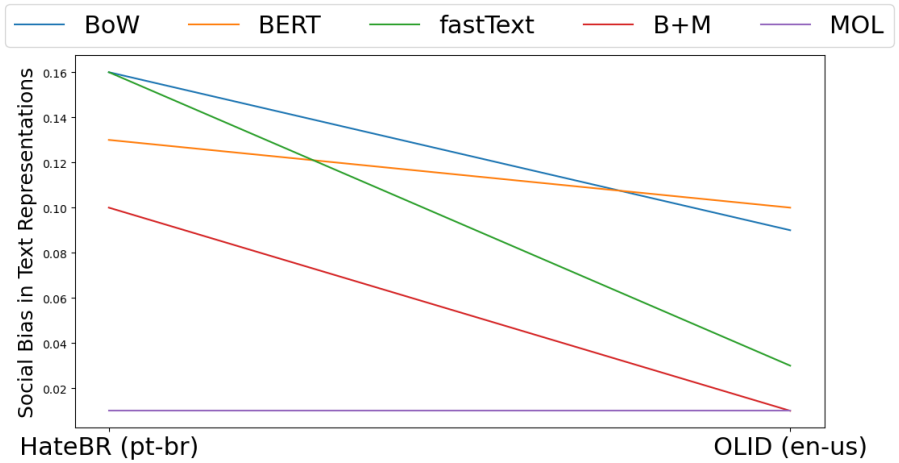


Figure: Distribution of social stereotypes bias in hate speech classifiers according to **SSA counterfactual explanations**.

HateBRXplain and Post-Hoc and Self-Explaining Methods (Portuguese).

Model [XAI method]	Plausibility				Faithfulness	
	IOU F1↑	Token Prec↑	Token Rec↑	Token F1↑	Comp.↑	Suff.↓
mBERT [LIME]	0.5828	0.7458	0.6936	0.6701	0.8809	0.0134
mBERT [SHAP]	0.6628	0.7143	0.7520	0.6897	0.9324	0.0172
BERTimbau [LIME]	0.5857	0.7557	0.6848	0.6698	0.9094	0.0237
BERTimbau [SHAP]	0.6600	0.7489	0.7099	0.6831	0.8458	0.0215
DistilBERTimbau [LIME]	0.6457	0.7614	0.7276	0.7003	0.9407	0.0115
DistilBERTimbau [SHAP]	0.6200	0.7543	0.6862	0.6720	**0.9475**	0.0114
PTTS [LIME]	0.6057	0.7487	0.6978	0.6776	0.5654	0.0016
PTTS [SHAP]	**0.7400**	0.7177	**0.8378**	0.7362	0.6160	0.0083
SRA (Ours, BERTimbau, $\alpha = 10$)	0.716 (±0.025)	**0.935** (±0.005)	0.668 (±0.014)	**0.745** (±0.010)	0.454 (±0.114)	-0.036 (±0.016)

Figure: Evaluation for **HateBRXplain (Portuguese) post-hoc and self-explanations** generated by LIME, SHAP, the proposed Supervised Rational Attention.

HateXplain and Post-Hoc and Self-Explaining Methods (English).

Model	Acc.↑	Macro F1↑	AUROC↑	GMB-Sub.↑	GMB-BPSN↑	GMB-BNSP↑	IOU F1↑	Token F1↑	AUPRC↑	Comp.↑	Suff.↓
CNN-GRU [LIME]	0.627	0.606	0.793	0.654	0.623	0.659	0.167	0.385	0.648	0.316	**-0.082**
BiRNN [LIME]	0.595	0.575	0.767	0.660	0.640	0.671	0.162	0.361	0.605	0.421	-0.051
BiRNN-Attn [Attn]	0.621	0.614	0.795	0.653	0.662	0.668	0.167	0.369	0.643	0.278	-0.001
BiRNN-Attn [LIME]	0.621	0.614	0.795	0.653	0.662	0.668	0.162	0.386	0.650	0.308	-0.075
BiRNN-HateXplain [Attn]	0.629	0.629	0.805	0.691	0.691	0.674	0.222	0.506	**0.841**	0.281	0.039
BiRNN-HateXplain [LIME]	0.629	0.629	0.805	0.691	0.691	0.674	0.174	0.407	0.685	0.343	-0.075
BERT [Attn]	0.690	0.674	0.843	0.762	0.709	0.757	0.130	0.497	0.778	0.447	0.057
BERT [LIME]	0.690	0.674	0.843	0.762	0.709	0.757	0.118	0.468	0.747	0.436	0.008
BERT-HateXplain [Attn]	**0.698**	**0.687**	0.851	0.807	0.745	0.763	0.120	0.411	0.626	0.424	0.160
BERT-HateXplain [LIME]	**0.698**	**0.687**	0.851	0.807	0.745	0.763	0.112	0.452	0.722	**0.500**	0.004
SRA (Ours)	0.696 (±0.007)	0.682 (±0.010)	**0.855** (±0.002)	**0.850** (±0.001)	**0.817** (±0.005)	**0.891** (±0.004)	**0.539** (±0.005)	**0.651** (±0.002)	0.753 (±0.001)	0.417 (±0.019)	-0.013 (±0.012)

Figure: Evaluation for **HateXplain (English) post-hoc and self-explanations** generated by LIME, SHAP, the proposed Supervised Rational Attention.

Post-Hoc Explanations for Automated Fact-Checking

Methods	Plausibility				Faithfulness	
	IOU F1 ↑	Token Precision ↑	Token Recall ↑	Token F1 ↑	Comp. ↑	Suff. ↓
mBERT-LIME	0.1098	**0.4378**	0.3913	0.3698	**0.2961**	**-0.0546**
mBERT-SHAP	**0.1529**	0.4312	**0.5111**	**0.4285**	0.2868	-0.0491

Table: Evaluation of **SELFAR post-hoc explanations** generated by LIME and SHAP methods.

Method	Readability			
	Flesch Reading Ease		Szigriszt Pazos Index	
Zero Shot Prompts	True	Fake	True	Fake
	0.77	**0.84**	-1519.48	**-1361.47**

Table: Evaluation of **SELFAR post-hoc explanations** generated from zero-shot prompts.

Conclusions

- My research tackles two relevant challenges: **detecting fake news and hate speech**, but doing so in a way that is **transparent, fair, and socially responsible**.
- I developed artificial intelligence methods that are not **black boxes**, in other words, Artificial Intelligence (AI) models that can **explain** why they classify a message as false or offensive, **mitigating the risks of biases**.
- The goal is for these solutions to be used not only in academic settings, creating practical tools as **BrasilSemÓdio**, and unique resources for **Brazilian Portuguese** and **Hausa** low-resourced languages.
- Proposing transparency, trust, and responsibility in the use of AI, with real impact on a **safe, factual and fair digital environment**. Of course, there are still challenges, but the results so far show that **explainable AI and fairness are both feasible and essential**.

Research Impact

1. 15 published papers in top-tier international NLP conferences (10 Qualis A1, 5 Qualis A3).
2. Over 300 citations, including references from prestigious institutions such as Harvard, the University of Michigan, and Carnegie Mellon.
3. Microsoft leveraged our HateBR dataset to train an LLM.
4. National and International Awards & Grants (e.g., Google LARA, ACL Grants, Brazilian Computer Society (SBC) Thesis and Dissertation Award (CTD) finalist.
5. Invited international visiting researcher at University of Southern California (USA) and Leibniz Institute (Germany).
6. Program committee member of top-tier NLP conferences/journals (e.g., EMNLP, ACL, NAACL, LREC, COLING, ICWSM, CIKM).
7. Co-organizer of top-tier international conferences/workshops (e.g., ICWSM, WOAH, DeepXplain).
8. Collaborations with researchers from 10 countries across 5 continents.
9. Inspiring +10 Ph.D., M.Sc., and undergraduate projects in Brazil.
10. Co-advisor of a master's student with publication in a top-tier NLP venue (e.g., DCC-UFMG).
11. Research outputs applicable for patents and registered systems with copyrights.

Acknowledgments

- **Parents and Friends**: my mother, father, sister and my pets; Isabelle Carvalho, Ilson Diniz, and Danilo Reis.

- **Professors**: Ricardo Marcacini (ICMC), Roseli Romero (ICMC), Adenilso Simao (ICMC), Thiago Pardo (ICMC), Fabricio Benevenuto (UFMG), Mirella Moro (UFMG), Morteza Dehghani (University of Southern California), Eduard Hovy (Carnegie Mellon University).

- **Collaborators**: Jackson Trager (University of Southern California), Zohar Rabinovich (University of Southern California), Flor Miriam Plaza-del-Arco (Leiden University), Surendrabikram Thapa (Virginia Polytechnic Institute), Ameeta Agrawal (Portland State University), Ibrahim Ahmad (Northeastern University), Diego Alves (Saarland University), Shamsuddeen Muhammad (Imperial College London), Matteo Guida (University of Melbourne), Kokil Jaidka (National University of Singapore), Marcello Gecchele (Tokyo Institute of Technology), Ali Hürriyetoğlu (KNAW Humanities), Wolfgang Schmeisser (University of Barcelona), Idris Abdulmumin (University of Pretoria, South Africa), Diallo Mohamed (University of Saint Thomas Aquinas), Isadora Salles (UFMG), Samuel Guimarães (UFMG), and Fabiana Góes (ICMC).

- **Grants**: Google, Fapesp, and Capes.

Thank you!

franciellealvargas@gmail.com

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*