Para Além da Medida de Toxicidade: Um Passo em Direção à Compreensão do Viés em Nível de Token

Lorenzo Puppi Vecchi | Emerson Cabrera Paraiso

Este estudo se concentra na análise da toxicidade na comunicação online, com ênfase nos vieses textuais em contextos tóxicos e racistas.

Problema

O problema central abordado é a crescente preocupação com a toxicidade na comunicação online e sua relevância na criação de ambientes inclusivos e respeitosos na internet. Especificamente, investigamos os vieses textuais que contribuem para a toxicidade e o racismo nas interações online.

Objetivo

O objetivo deste trabalho é compreender os vieses subjacentes e os mecanismos que levam à toxicidade textual em contextos online, classificando conteúdos tóxicos e identificando as partes das sentenças que contribuem para cada tipo de viés. Buscamos promover a conscientização sobre a importância de combater a toxicidade na comunicação online e fomentar ambientes virtuais mais inclusivos e respeitosos.

Introdução

Este estudo se concentra na análise da toxicidade na comunicação online, com ênfase nos vieses textuais em contextos tóxicos e racistas.

Problema

O problema central abordado é a crescente preocupação com a toxicidade na comunicação online e sua relevância na criação de ambientes inclusivos e respeitosos na internet. Especificamente, investigamos os vieses textuais que contribuem para a toxicidade e o racismo nas interações online.

Objetivo

O objetivo deste trabalho é compreender os vieses subjacentes e os mecanismos que levam à toxicidade textual em contextos online, classificando conteúdos tóxicos e identificando as partes das sentenças que contribuem para cada tipo de viés. Buscamos promover a conscientização sobre a importância de combater a toxicidade na comunicação online e fomentar ambientes virtuais mais inclusivos e respeitosos.

Classificação binária

XAI

Introdução

PUCPR
GRUPO MARISTA

Modelos de linguagem pré-treinados (self-supervised learning)
recentes alcançaram melhorias em compreensão de linguagem natural. Estes usualmente:

- Treinam com base na similaridade de palavras e frases.
- O objetivo de otimização maximiza a probabilidade das corpora de treinamento.
- A coerência de palavras e frases frequentemente usadas juntas aumenta com o modelo treinado.
- Corpora de treinamento são gerados por seres humanos e podem conter viés social e estereótipos, incluindo gênero, raça e religião.

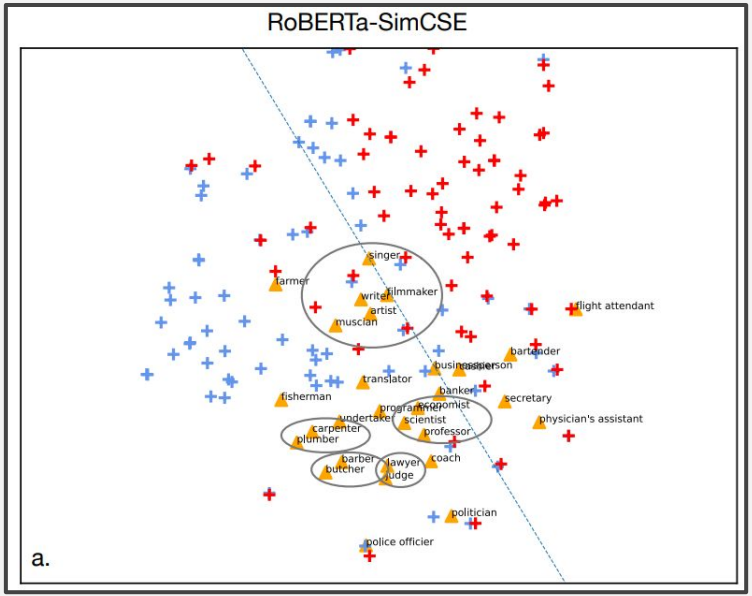
Masked Training

Autoregressive

Trabalhos anteriores

SimCSE: Simple Contrastive Learning of Sentence Embeddings.
Princeton University

Logic Against Bias: Textual Entailment Mitigates Stereotypical Sentence Reasoning.
MIT Computer Science and Artificial Intelligence Laboratory



Substantivos profissionais

Termos femininos

Termos masculinos

Trabalhos anteriores

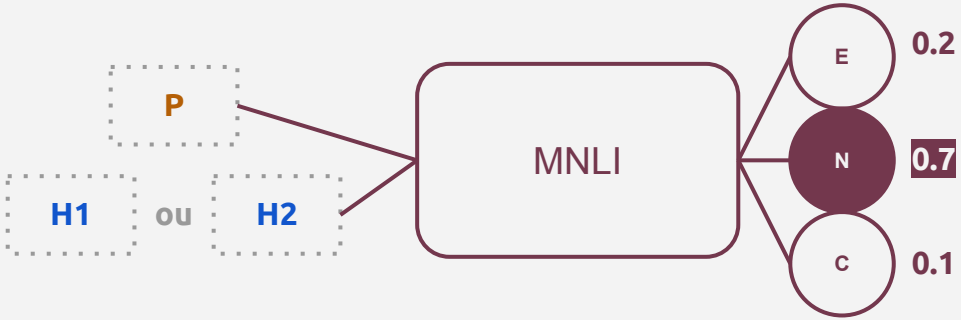
O aprendizado por inferência textual *(MNLI/NLI - Natural language inference)* enfatiza a lógica em vez da similaridade semântica:

- A inferência textual, não requer uma inferência lógica estrita; em vez disso, envolve uma premissa verdadeira que torna a hipótese provavelmente verdadeira.
- A contradição ocorre quando a premissa verdadeira torna a hipótese provavelmente falsa.
- Uma sentença pode ser inferida, neutra ou contraditória em relação a sentenças semelhantes ou não semelhantes semanticamente.
- Modelos de inferência textual são menos propensos a realizar raciocínio estereotipado baseado na similaridade de texto.

P Esta pessoa trabalha com medicina

H1 Esta pessoa é mulher

H2 Esta pessoa é homem

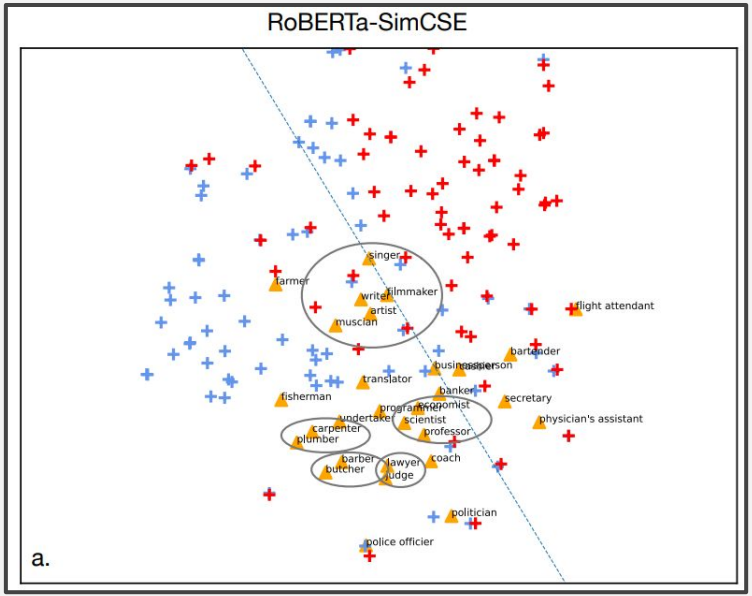


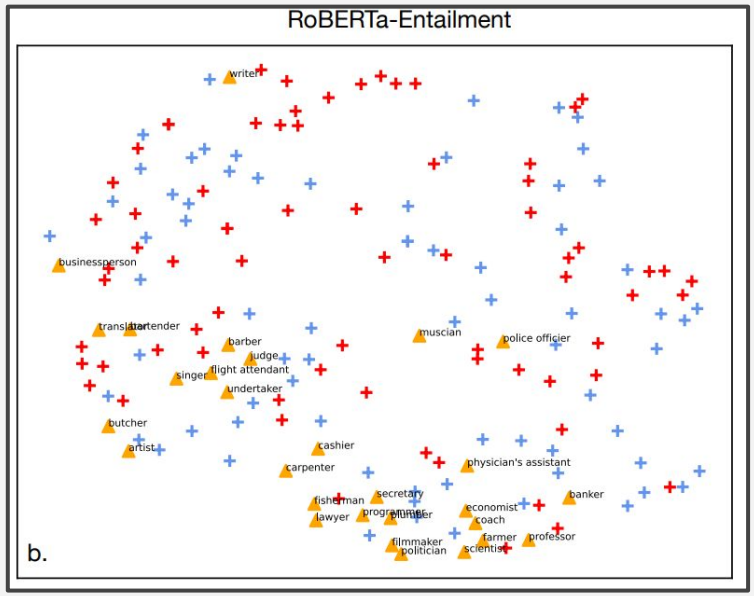
e = envolvimento | **n** = neutro | **c** = contradição

Trabalhos anteriores

SimCSE: Simple Contrastive Learning of Sentence Embeddings.
Princeton University

Logic Against Bias: Textual Entailment Mitigates Stereotypical Sentence Reasoning.
MIT Computer Science and Artificial Intelligence Laboratory





Substantivos profissionais

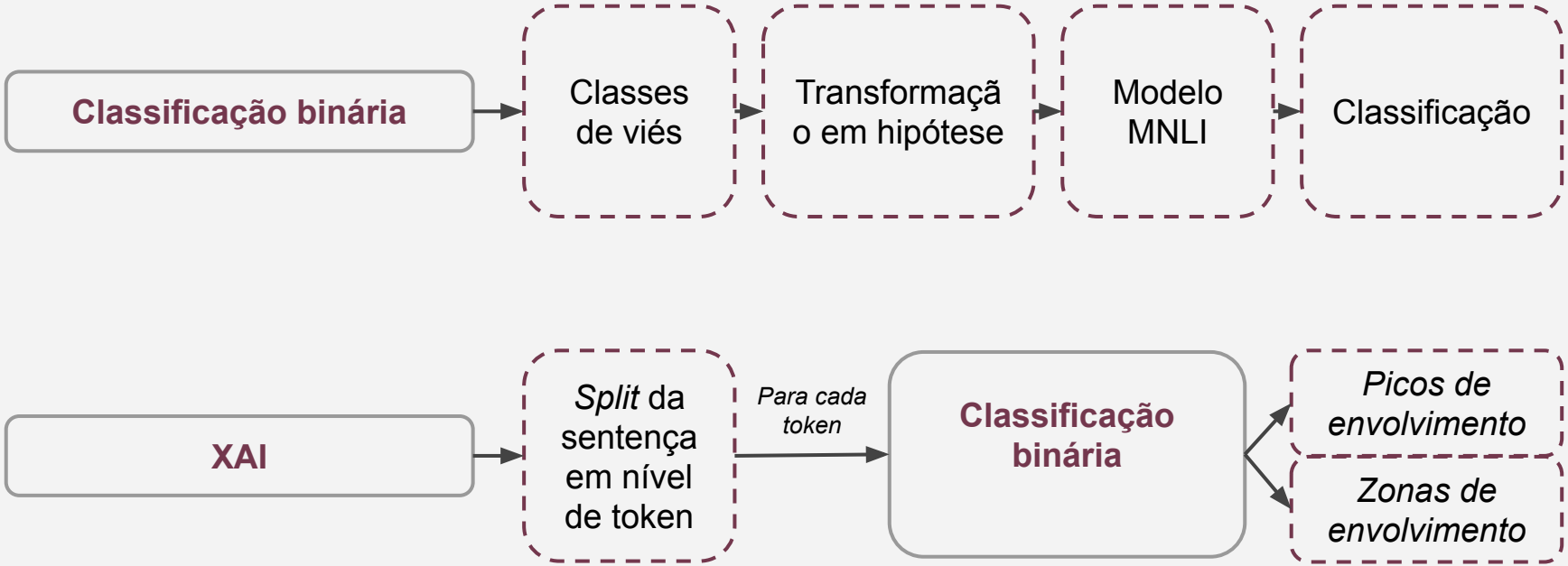
Termos femininos

Termos masculinos

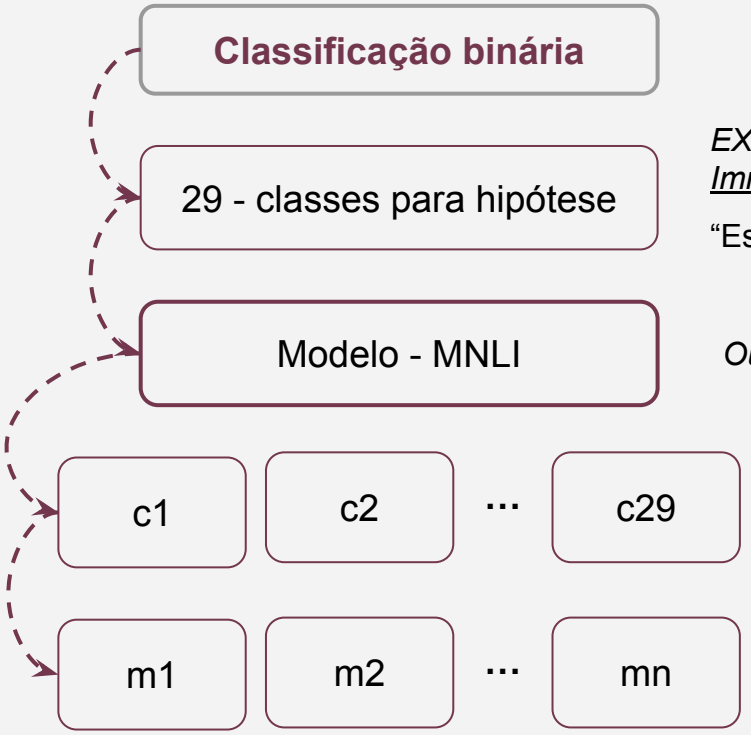
Trabalhos anteriores

PUCPR
GRUPO MARISTA

Método


Classificação binária
Classes de viés
Transformaçã o em hipótese
Modelo MNLI
Classificação
XAI
Split da sentença em nível de token
Para cada token
Classificação binária
Picos de envolvimento
Zonas de envolvimento

Método



EX:
Immigration status (Immigrants)

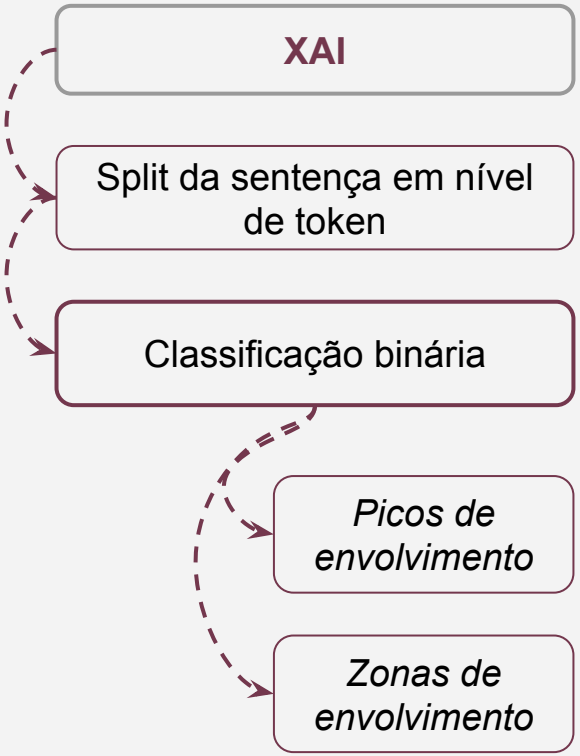
"Esta sentença fala sobre imigrantes"

Output "Entailment" utilizado.

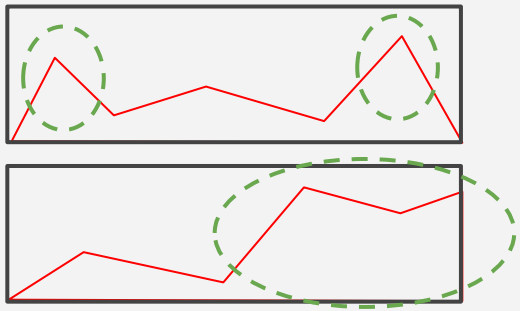
Cada classe se torna uma característica

Múltiplos modelos de machine learning são adotados para, utilizando estas características, aprender a classificar

Método



Output "Entailment" utilizado.



Método

Resultados

Comparação em diferentes splits de treino e testes com o trabalho:
"Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection".
Facebook AI Research

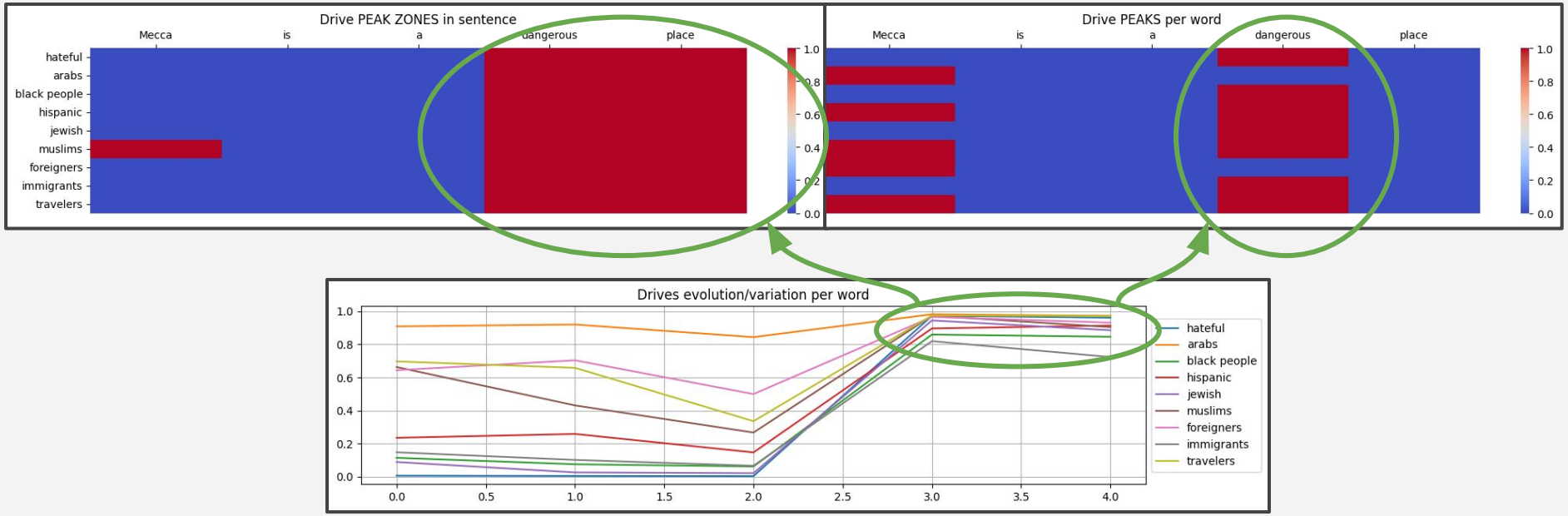
A comparação foi feita entre algoritmos simples de machine learning que utilizam caracteristicas de um modelo de NLI, vs um Roberta-base model de 125M de parâmetros, treinado em 8 GPUs.

Train	Test F1 Macro	1	2	3	4
1	OURS	0.779	0.741	0.617	0.637
	PREV	**0.922**	0.629	0.477	0.524
2	OURS	0.74	0.736	0.702	0.701
	PREV	0.807	0.765	**0.774**	**0.749**
3	OURS	0.724	0.74	0.66	0.696
	PREV	0.727	**0.785**	0.741	0.732
4	OURS	0.7	0.733	0.695	0.686
	PREV	0.723	0.768	0.772	0.696
1+2	OURS	0.785	0.75	0.673	0.684
	PREV	0.911	0.747	0.747	0.716
1+2+3	OURS	0.78	0.758	0.672	0.695
	PREV	**0.912**	0.77	0.746	**0.739**
1+2+3+4	OURS	0.776	0.757	0.673	0.683
	PREV	0.903	**0.779**	**0.768**	0.729

Train	Test	
1	1	PREV
	2	OURS: SVM 0.741>0.629
	3	OURS: SVM 0.617>0.477
	4	OURS: SVM 0.637>0.524
2	1	PREV
	2	Competitive: < than 5pp - 0.736<0.765
	3	PREV
	4	Competitive: < than 5pp - 0.701<0.749
3	1	Competitive: < than 5pp - 0.724<0.727
	2	Competitive: < than 5pp - 0.74<0.785
	3	PREV
	4	Competitive: < than 5pp - 0.696<0.732
4	1	Competitive: < than 5pp - 0.7<0.723
	2	Competitive: < than 5pp - 0.733<0.768
	3	PREV
	4	Competitive: < than 5pp - 0.686<0.696
1+2	1	PREV
	2	OURS: SVM 0.75>0.747
	3	PREV
	4	Competitive: < than 5pp - 0.684<0.716
1+2+3	1	PREV
	2	Competitive: < than 5pp - 0.758<0.77
	3	PREV
	4	Competitive: < than 5pp - 0.695<0.739
1+2+3+4	1	PREV
	2	Competitive: < than 5pp - 0.757<0.779
	3	PREV
	4	Competitive: < than 5pp - 0.683<0.729

Resultados - Classificação

“Mecca is a dangerous place”



https://article-2023.vercel.app/

Resultados - XAI

PUCPR
GRUPO MARISTA

Gostaria de agradecer a todos por dedicarem seu tempo para participar desta apresentação. A análise da toxicidade na comunicação online é um tema de extrema importância, pois influencia diretamente na criação de ambientes virtuais inclusivos e respeitosos.

Nossa pesquisa, apesar de inicial, busca compreender os vieses textuais que contribuem para a toxicidade e o racismo nas interações online, visando promover a busca por soluções para tornar a internet um lugar mais acolhedor para todos, sempre respeitando a liberdade de expressão.

Queremos agradecer à CAPES Brasil (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) por apoiar parcialmente este trabalho.

Obrigado!

PUCPR
GRUPO MARISTA

CAPES

Para Além da Medida de Toxicidade: Um Passo em Direção à Compreensão do Viés em Nível de Token

Lorenzo Puppi Vecchi | Emerson Cabrera Paraiso

lorenzo.vecchi@ppgia.pucpr.br

paraiso@ppgia.pucpr.br