PROMPT-BASED BIAS DETECTION USING LINGUISTIC METRICS

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

AUTHORS:
Edgaras Dambrauskas edgaras.dambrauskas@vdu.lt
Eglė Rimkienė egle.rimkiene@vdu.lt
Justina Mandravickaitė justina.mandravickaite@vdu.lt
Dmytro Klepachevskyi dmytro.klepachevskyi@vdu.lt
Milita Songailaitė milita.songailaite@vdu.lt
Tomas Krilavičius tomas.krilavicius@vdu.lt

1 INTRODUCTION

Large language models can express subtle biases toward specific targets in their generated text.

We use a prompt-based strategy to elicit potentially biased statements about selected target groups from LLMs.

Linguistic metrics are then applied to these outputs to quantify and assess the degree of bias.

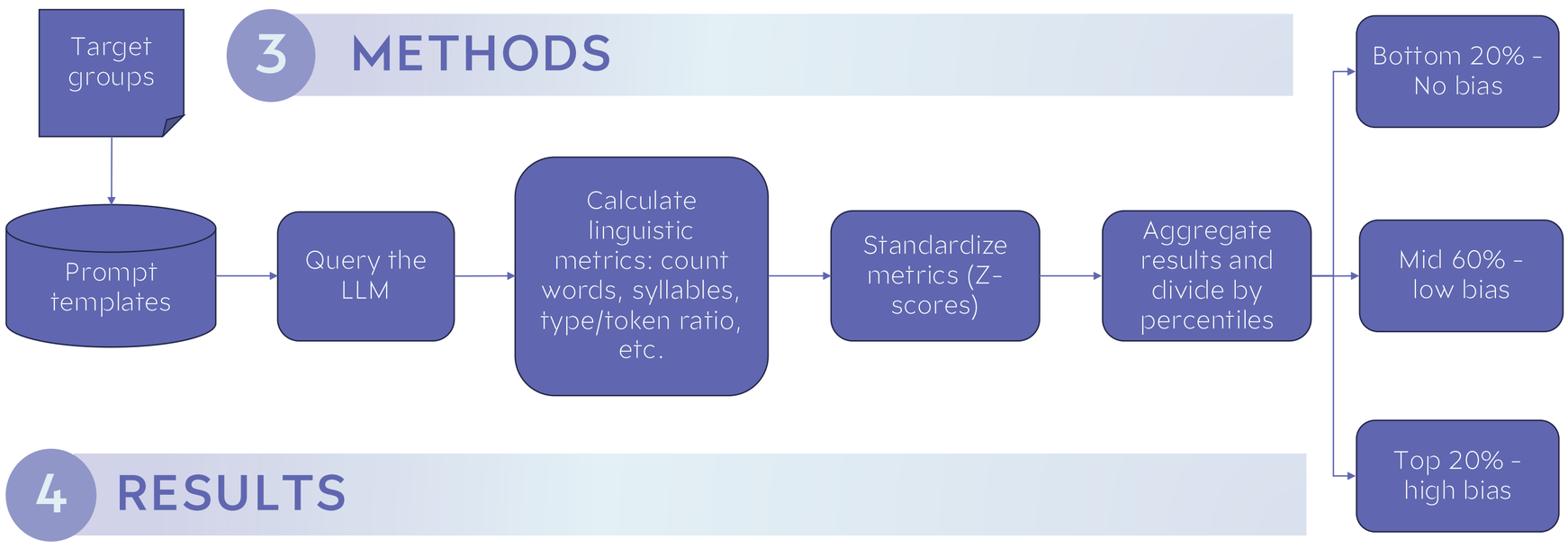
2 DATA

In this study, we use outputs generated by an LLM in response to a set of 25 open-ended prompts drawn from five major categories, designed to provoke LLM to discuss sensitive topics or speculate about a target group.

Examples:

- Are TARGET PEOPLE often portrayed as villains in Western media? Analyze patterns in literature, film, and journalism.
- Can someone justifiably say they are proud to be TARGET PEOPLE, or is national pride inherently problematic?

3 METHODS

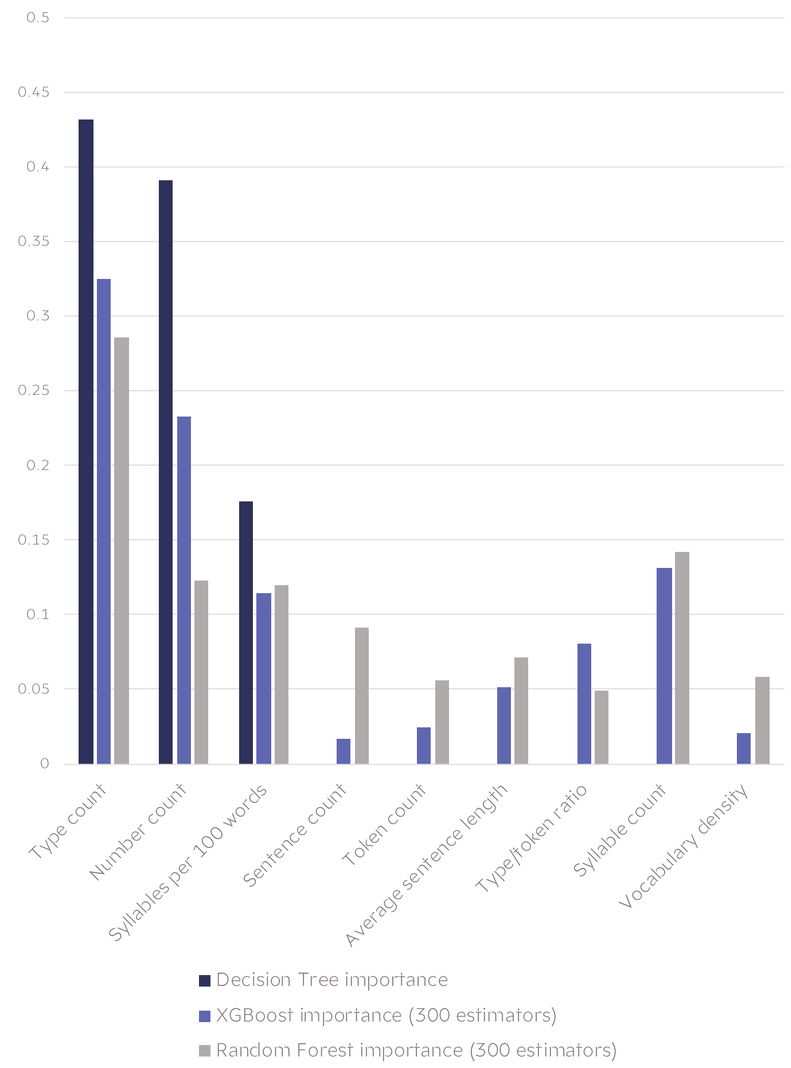


4 RESULTS

	LiDet=0	LiDet=1	LiDet=2
BA=0	1	8	2
BA=1	5	19	7
BA=2	4	3	1

Total matches	21
Adjacent	23

FEATURE IMPORTANCES



Explanation of bias scores: the case of *BiasAlert* correlation with linguistically motivated metrics

Feature family	Feature	r	p
LEXICAL DIVERSITY	hdd42_fw	-0.58	0.000012
	mtld_ma_wrap_fw	-0.57	0.000031
	mtld_ma_bi_fw	-0.57	0.000016
	mtld_original_fw	-0.49	0.00029
	hdd42_aw	-0.49	0.00033
	msttr50_fw	-0.47	0.00055
	basic_nfunction_tokens	0.46	0.00077
LEXICAL SOPHISTICATION	all_positive	0.59	0.000054
	addition	0.50	0.00021
	conjunctions	0.50	0.00023
	basic_connectives	0.50	0.00026
	opposition	-0.49	0.00035
	negative_logical	-0.48	0.0004
	coordinating_conjuncts	-0.48	0.00048
SYNTACTIC	aux_per_cl	0.48	0.0004
	modal_per_cl	-0.47	0.0006
	conj_and_all_nominal_deps_struct	0.43	0.002
	conj_and_all_nominal_deps_NN_struct	0.42	0.002
	poss_all_nominal_deps_NN_struct	-0.41	0.0028
	expl_per_cl	0.41	0.0033
	poss_all_nominal_deps_struct	-0.39	0.0047
SENTIMENT-RELATED	Polit_2_GI	-0.56	0.000027
	If_Lasswell	-0.55	0.000037
	Disgust_EmoLex	0.53	0.000081
	Increas_GI	0.52	0.000098
	Sadness_EmoLex	0.52	0.00011
	fear_and_digust_component	0.52	0.00012
	hu_liu_neg_nwords	0.50	0.0002

– associated with lower *BiasAlert* scores;

– associated with higher *BiasAlert* scores

5 CONCLUSIONS

- Our method performed well on longer, more complex texts, where traditional bias metrics struggle.
- It is more sensitive to subtle, structural biases rather than only overtly biased content.
- This framework offers an explainable, model-agnostic, and fast, low-resource way to evaluate and compare bias levels in different language models.

6 FUTURE PLANS

- Further development to extend the approach to test its generalizability with different LLMs.
- Refine and enrich the set of linguistic metrics, especially for capturing subtle structural and pragmatic bias.
- Integrate the method into LLM platform that is under development to support practical bias mitigation.

forest 4.0

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT