Examining the use of Large Language Models (LLMs) for policymaking in South Africa

Evaluating contextual bias in climate policy recommendations generated by LLMs

INTRO

- LLMs have the potential in policymaking, but biases in training data present challenges.
- This project uses multiple LLMs and machine learning to address the biases.

METHODS

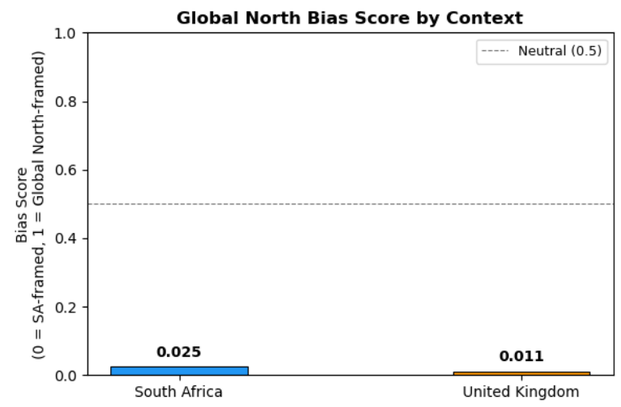
1. N = 80 newspaper articles.
2. Collected climate change articles from 4 online news outlets (i.e., Business Day, Mail & Guardian, Business Live, The Independent)
3. EDA: LDA topic modelling (10 topics), TextBlob sentiment analysis, word frequency plots on pre-processed corpus
4. Generated policy recommendations from 3 LLMs (Cohere Aya, Meta Llama, and Mistral) from scenario-based prompts.
5. Modelling and evaluation:
   - Bias Analysis:
     - cosine similarity, LDA topic alignment scoring, and bias score
   - Classification: Logistic regression to predict SA vs. UK LLM output context
     - Accuracy score, recall, and F1-score
6. PCS Validation
   - Predictability: same prompts, 3 LLMs
   - Computability: Metrics used
   - Stability: LDA stability across 3 random seeds (Jaccard 0.0–0.43)

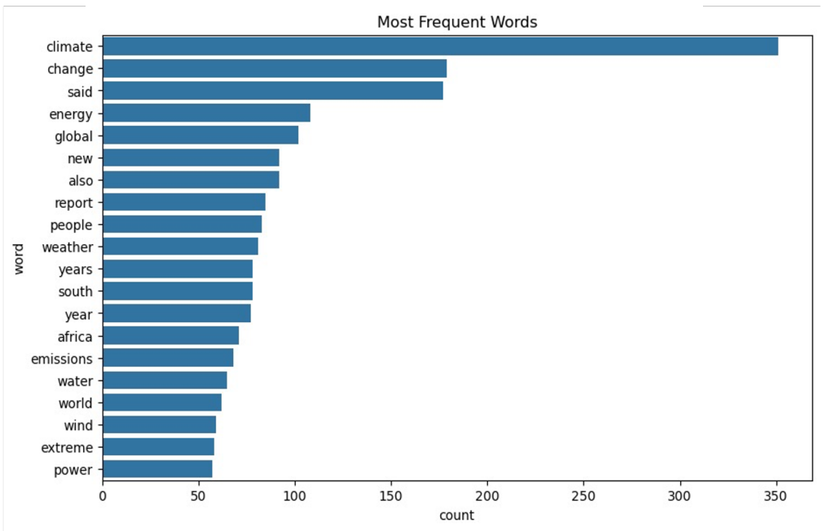
RESULTS

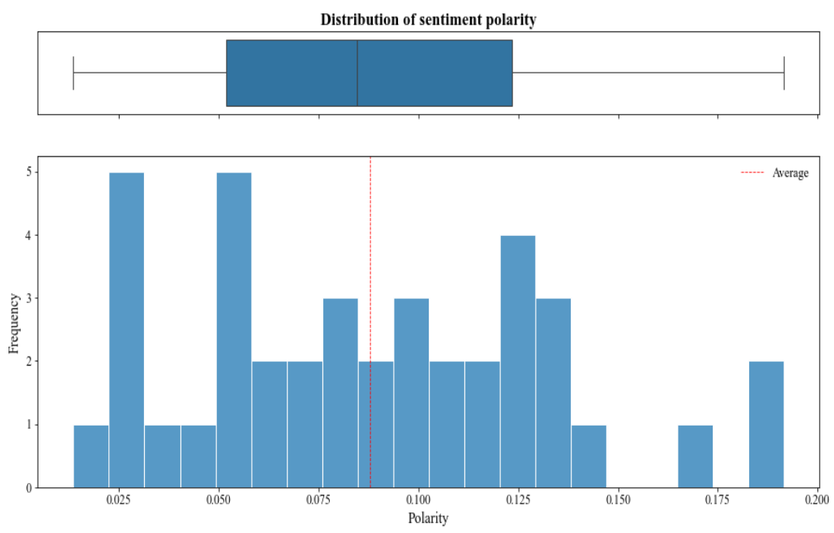
- Context sensitivity confirmed: all 3 LLMs scored higher SA topic alignment when given SA articles vs. UK articles (topic alignment delta positive across all models & scenarios).
- Mistral = most context-sensitive
- Llama = most generic
- Cohere = intermediate
- The model achieved an accuracy of 46.7%, with a strong recall for class 0 (South Africa); however, the model failed to identify instances of class 1 (United Kingdom) accurately, and resulted in an F1-score of 0.3

DISCUSSION

- The findings highlight the challenges of generating region-specific policy recommendations due to limited data and variations in climate change and policy lexicon.
- Training origin does not reliably predict regional adaptability.
- Keyword scoring only captures explicit terminology, and subtler biases go undetected.
- Limitations: Small corpus (80 articles) constrains LDA stability and classifier accuracy.
- Future research should increase the volume and variety of data sources and develop more rigorous evaluation frameworks for assessing the feasibility of LLM-generated policy recommendations to improve the classifier







Leeroy Mabena, Thandoluhle Moyo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me