Speaker Landscapes - Machine Learning Opens a Window on the Everyday Language of Opinion

Maria Schuld[1,2] Kevin Durrheim[2] Martin Mafunda[1]
[1] University of KwaZulu-Natal, [2] University of Johannesburg

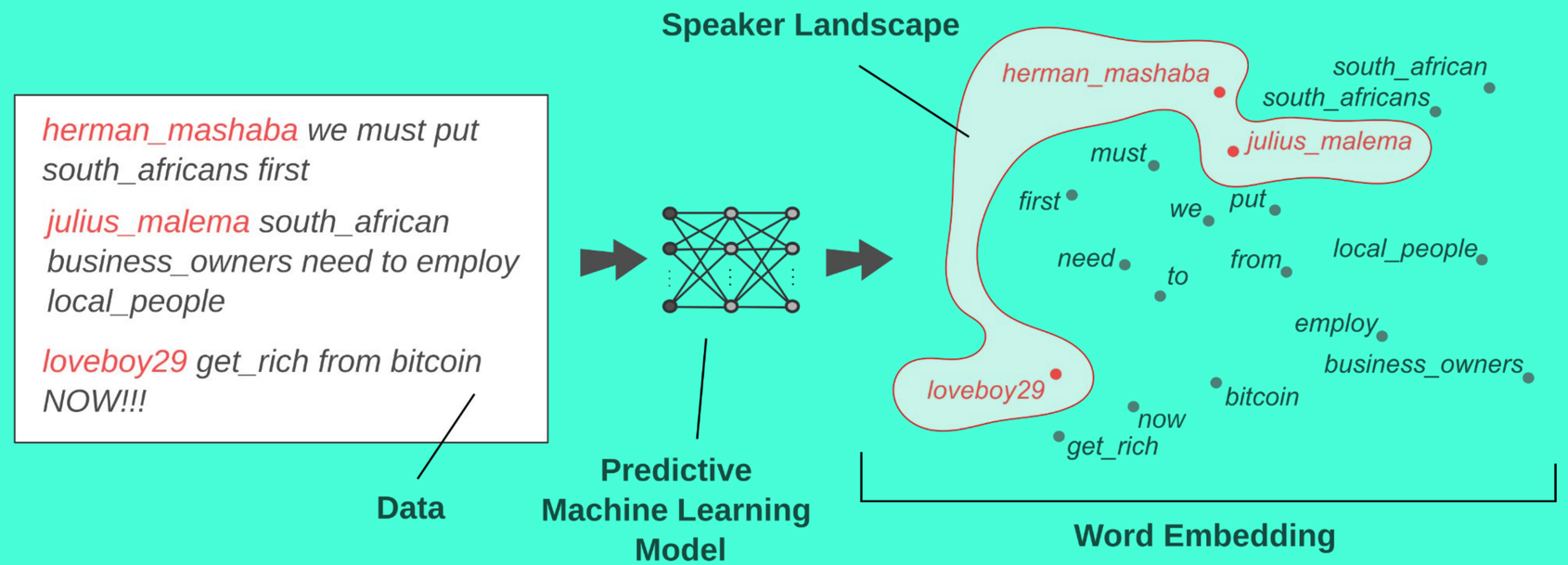
check out the pre-print for many more experiments

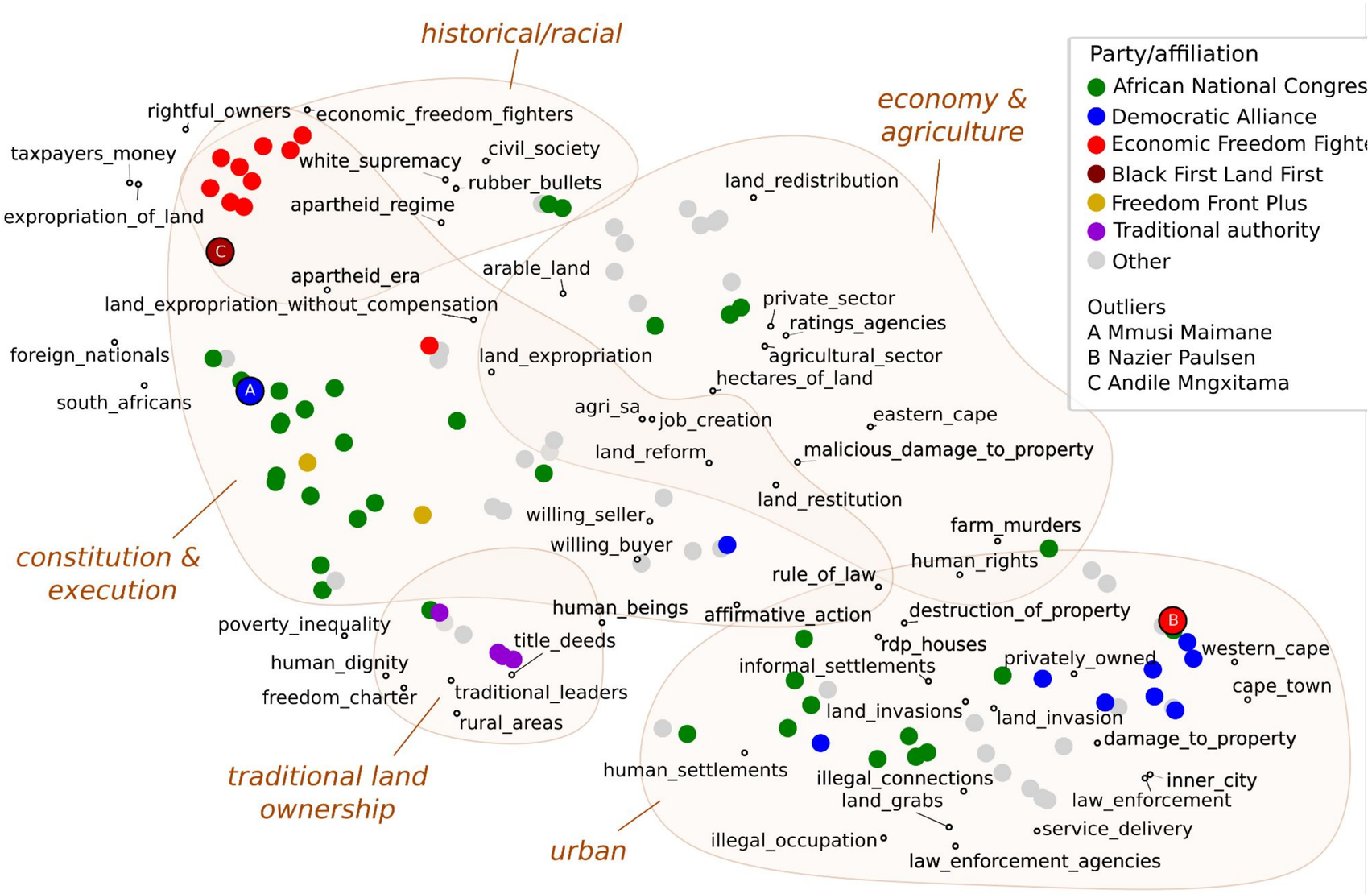
Idea

We train a word embedding [1] on speech data extended by a token representing the speaker.

The resulting speaker-token embedding, which we call “speaker landscape” [2], positions speakers who speak similarly closely together.

We want to find out if this model can help us to investigate political polarisation.





Case study

We trained a speaker landscape using quotes extracted from Media Monitoring Africa’s Southern African database of digitized print and online news media (2013- 2021).

Pretraining was performed with ~3 Mio general quotes, after which we fine-tuned the model with 3356 quotes related to the land debate in South Africa.

The plot on the left shows a 2-dimensional visualisation of the landscape. Each dot represents an embedded speaker-token. N-grams from the word embedding were added to give context.

Observations & Conclusions

While speaker landscapes are solely measuring how similarly individuals speak, they expose complex social patterns. For example, the above case study suggests that:

- *Speech similarity correlates well with political affiliation.*
  The landscape clearly distinguishes speakers by their party. As expected, South African opposition parties (Democratic Alliance, Economic Freedom Fighters, Freedom Front Plus) as well as the traditional authorities appear spatially clustered, while the ruling party (African National Congress) is spread out.
- *Speech similarity carries a signature of polarisation.*
  The Democratic Alliance and Economic Freedom Fighters, whose political dispositions are very distinct – pro-capital versus pro-poor respectively – are placed at the polar ends of the landscape.
- *Speech similarity reflects political allies and outliers.*
  For example, Mmusi Maimane who left the Democratic Alliance in 2019 is located far away from the blue cluster; Nazier Paulsen, an Economic Freedom Fighters member in the Western Cape parliament shares the language of the ruling Democratic Alliance; and Andile Mngxitama who was previously part of the Economic Freedom Fighters is still closely positioned with them.

We reproduced similar results in other case studies [3, 4], suggesting that speaker landscapes are a promising tool for social science research.

[1] Mikolov, Chen, Corrado, Dean (2013). Efficient estimation of word representations in vector space. International Conference on Learning Representations.
[2] Schuld, Durrheim, Mafunda (2023) Speaker landscapes: Machine learning opens a window on the everyday language of opinion. *Under submission.* [https://osf.io/smhn5]
[3] Sepahpour-Fard, Quayle, Schuld and Yasseri (2023) How does the audience affect the way we express our gender roles? *Under submission.* [arXiv:2303.12759v2]
[4] Durrheim, Schuld (2023) Group polarization on social media: Comparing the dynamics of interaction networks and language-based opinion distributions *Under submission.* [psyarxiv.com/n47xe/]