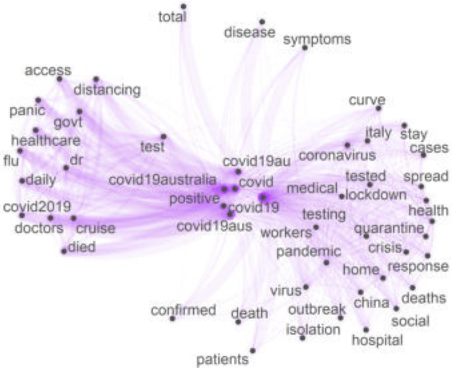
Text Mining the COVID-19 discourse in the Australian Twittersphere

Martin Schweinberger | Michael Haugh | Sam Hames



Why analyze COVID-19 discourse on Twitter?

- COVID-19 discourse treated as one big lump of words
- Focus on hashtags or use of small/unclean data sets

How can we as linguists improve text mining of discourse around COVID-19 on OzTwitter?

	Jan. - Apr. 2019		Jan. - Apr. 2020	
	Tweets	**Words/Elements**	**Tweets**	**Words/Elements**
Before processing	889,192	18,903,659	871,826	19,362,115
After processing	769,165	17,288,018	753,630	17,726,090
COVID-19 tweets			41,342	1,327,874

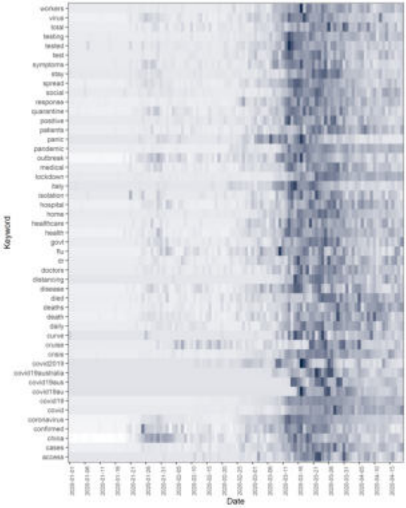


Figure 1: Heatmap of COVID19-related keywords

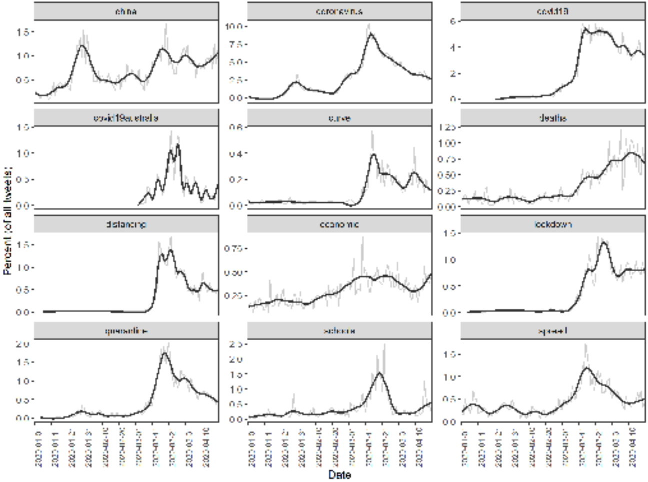


Figure 2: Linegraphs of selected COVID19-relates keyterms across time of COVID19-related keywords

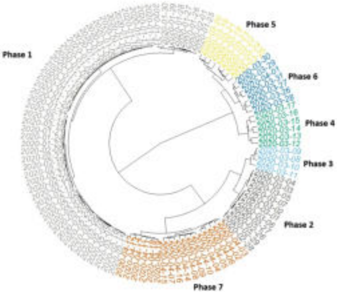


Figure 3: Results of the PAM clustering showing the data-driven periodization of the data

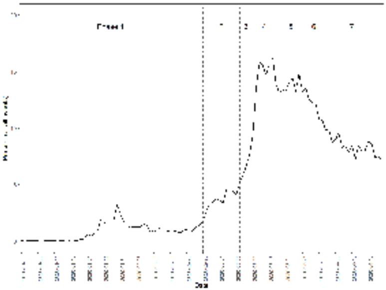


Figure 4: Percantages of COVID19-related tweets by period

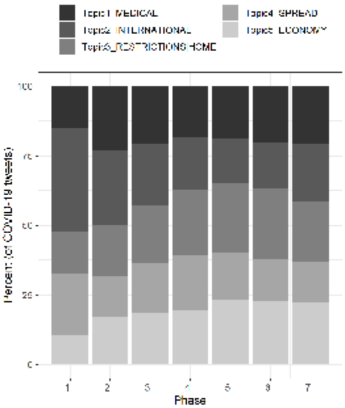


Figure 5: Distribution of topics across periods (bar plot)

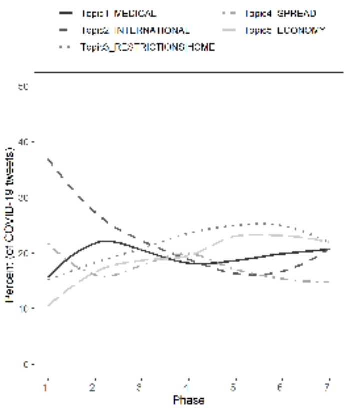


Figure 6: Distribution of topics across periods (loess smoothed)

Outlook

Aim: create a prototype of a text mining application that is both time-sensitive and differentiates between different sub-discourses (topics)

Moving forward

- Apply analysis to more data
- Apply same method to other phenomena (BLM, Bushfires)

Thank you very much!

Acknowledgements

We would like to thank the Digital Observatory Research Facility operated by the Institute for Future Environments at QUT who have compiled the data that the current study is based on.

Text Mining the COVID-19 discourse in the Australian Twittersphere

Martin Schweinberger | Michael Haugh | Sam Hames