Text Mining the COVID-19 discourse in the Australian Twittersphere

Martin Schweinberger (UQ)
Michael Haugh (UQ)
Sam Hames (QUT)

slides available at

www.martinschweinberger.de

m.schweinberger@uq.edu.au

R code upon request after embargo

Why analyze COVID-19 discourse on Twitter?

General

- Interesting topic
- Text mining: typically lacks periodization
- Text mining: one discourse rather than collection of discourses

Previous research

- COVID-19 discourse treated as an undifferentiated bag-of-words
- Focus on individual hashtags
- Small or unclean data sets (either only official channels/users | only few days are monitored)

Starting point

Research Question

Can we identify different phases in the COVID19 discourse on OzTwitter and, if so, how do these phases differ from each other (what characterizes each period)?

Hypothesis

We assume that COVID-19 discourse developed in these phases:

Phase 1: China $\rightarrow$ Phase 2: Covid/Lockdown $\rightarrow$ Phase 3: Economy

Data

What kind of Twitter data?

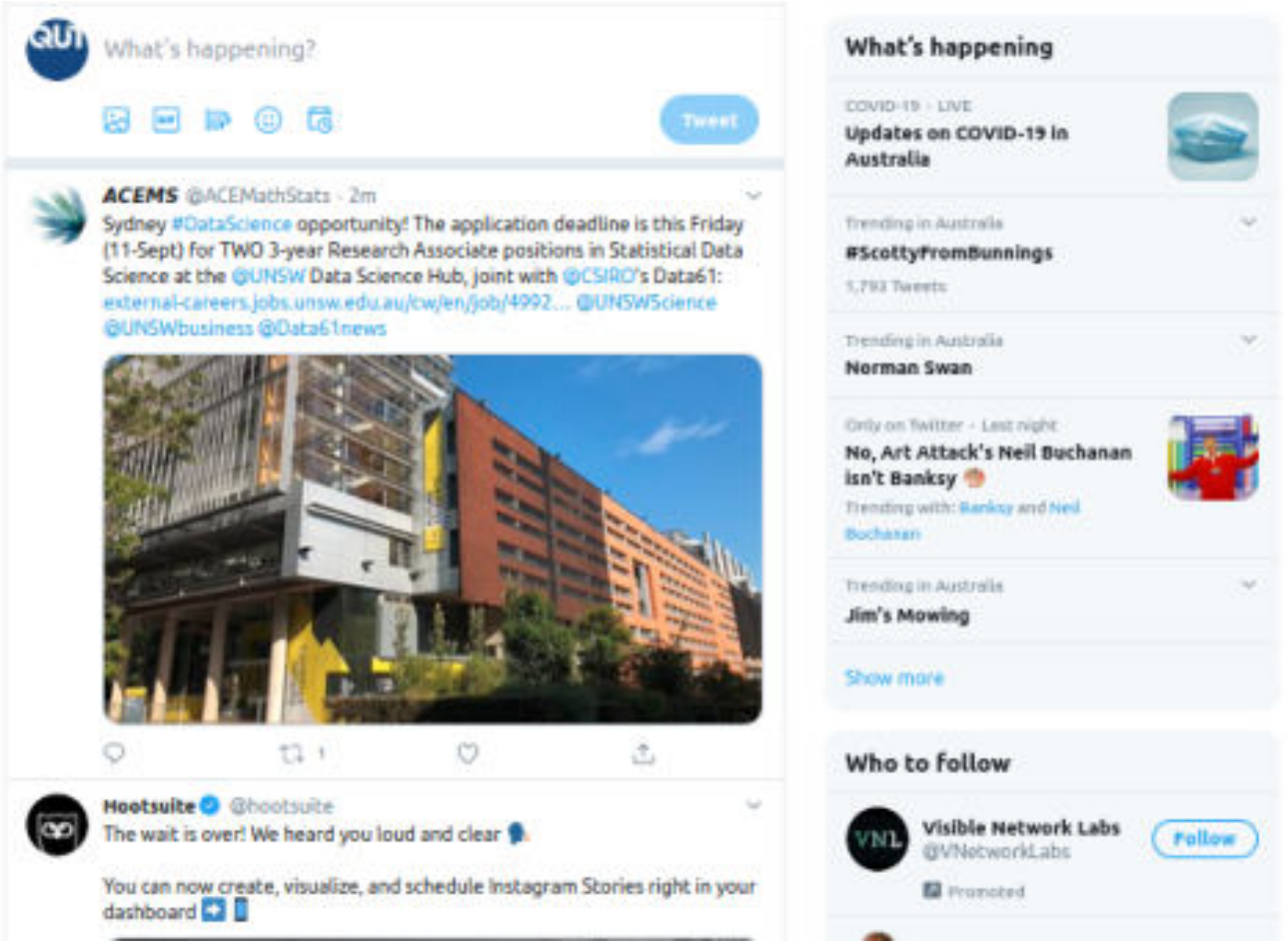

Figure 1: Twitter data before extraction

What kind of Twitter data?

Figure 2: User information

How can you collect Twitter data?

- (One of) the Official Twitter API's, optionally through a wrapper library like `https://github.com/DocNow/twarc`
- scraping or related tools like `https://github.com/twintproject/twint`

The Australian Twittersphere

- 500,000 accounts identified as Australian
- 100,000 daily active tweeters
- Longitudinal tweets collected since May 2018
- 25 million tweets/month

Twitter corpus

Entire Twitter corpus: app. 1.7 million tweets (38 million words/elements)

- Corpus 1
  1 percent sample of all Australian tweets from Jan 1 - Apr 15, 2019 (control)
- Corpus 2
  1 percent sample of all Australian tweets from Jan 1 - Apr 15, 2020

Idea: Create training set with non-COVID19 tweets from 2019 data and COVID19 tweets from the 2020 data
$\rightarrow$ Use training set to identify keywords and build a classifier

Workflow

Entire workflow in R (R Core Team 2020)

- Data processing|cleaning
- Data viz|analysis
  - Classification: identifying COVID19 tweets
  - Keyword extraction
  - Periodization
  - Topic modeling
  - Sentiment analysis

Data processing

Tweet-level

- Removal of tweets with non ASCII - elements (e.g. Chinese|Japanese|Korean characters)
- Removal of tweets with non-English language (Spanish: *corona* = *crown*!)
- Conversion to lower case

Word level

- Removal of stopwords
- No lemmatization!
- No spell checking/correcting!

Data summary

	2019		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

Analysis

Classification

Identifying tweets that are COVID-19 related

Problem

- COVID-19 related tweets may not mention COVID-19
- Analysis should be possible on notebook (no HPC)

Solution

- Support Vector Machine-based classifier (linear kernel)
- Training set: 5,000 tweets (750 COVID-19, 4,250 non-COVID19) $\rightarrow$ length of vector: 16,087
- Test set: 1,250 tweets $\rightarrow$ 100 % prediction accuracy

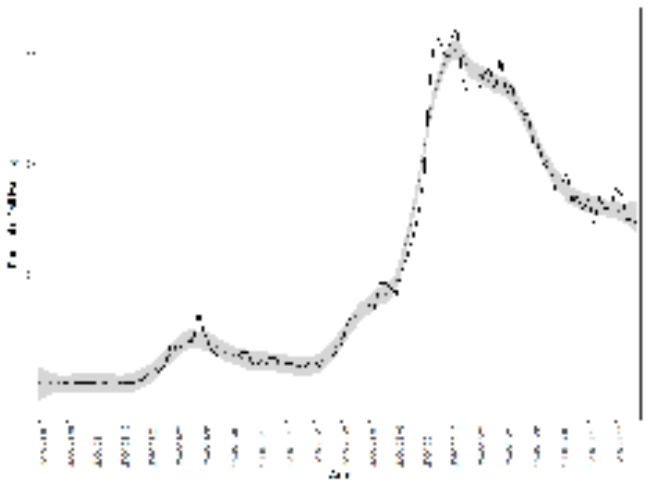

Figure 3: Percent of COVID19 tweets of all tweets per day

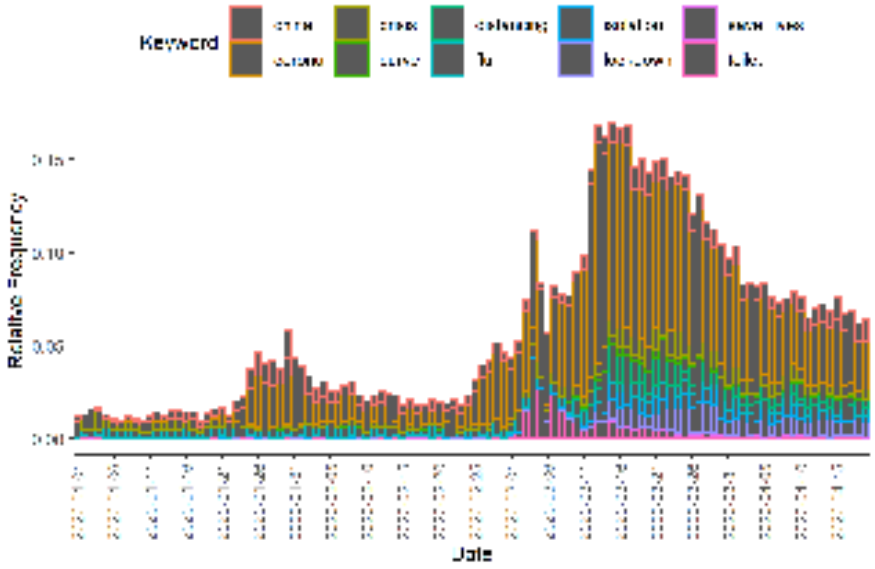

Figure 4: Percent of selected COVID19-related key terms per day

Keyword extraction

Identifying keywords that are significantly associated with COVID-19

Solution

- Fisher's Exact tests (with Benjamini-Hochberg correction for multiple/repeated tests)
- Applied to all word types in the data $\rightarrow$ 49 terms that are significantly and positively correlated with COVID-19 discourse

	N COVID19	**N non-COVID19**
Element	a	b
Other elements	c	d

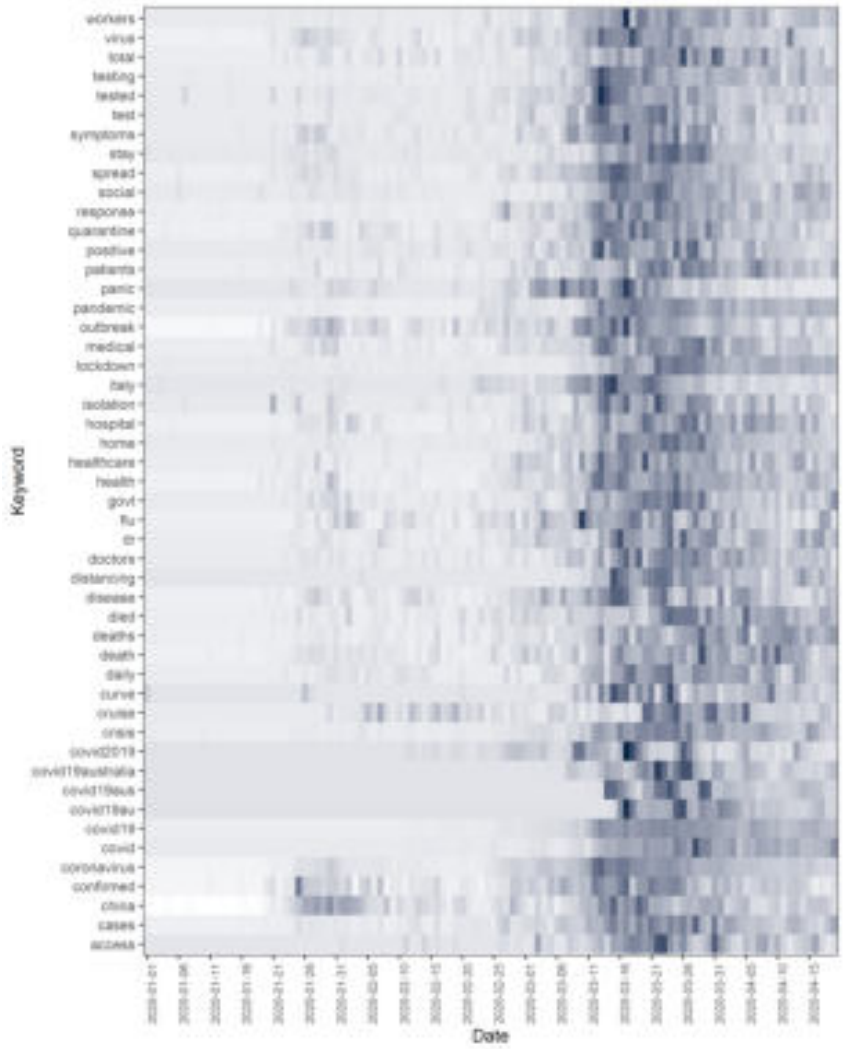

Figure 5: Heatmap of COVID19-related keywords

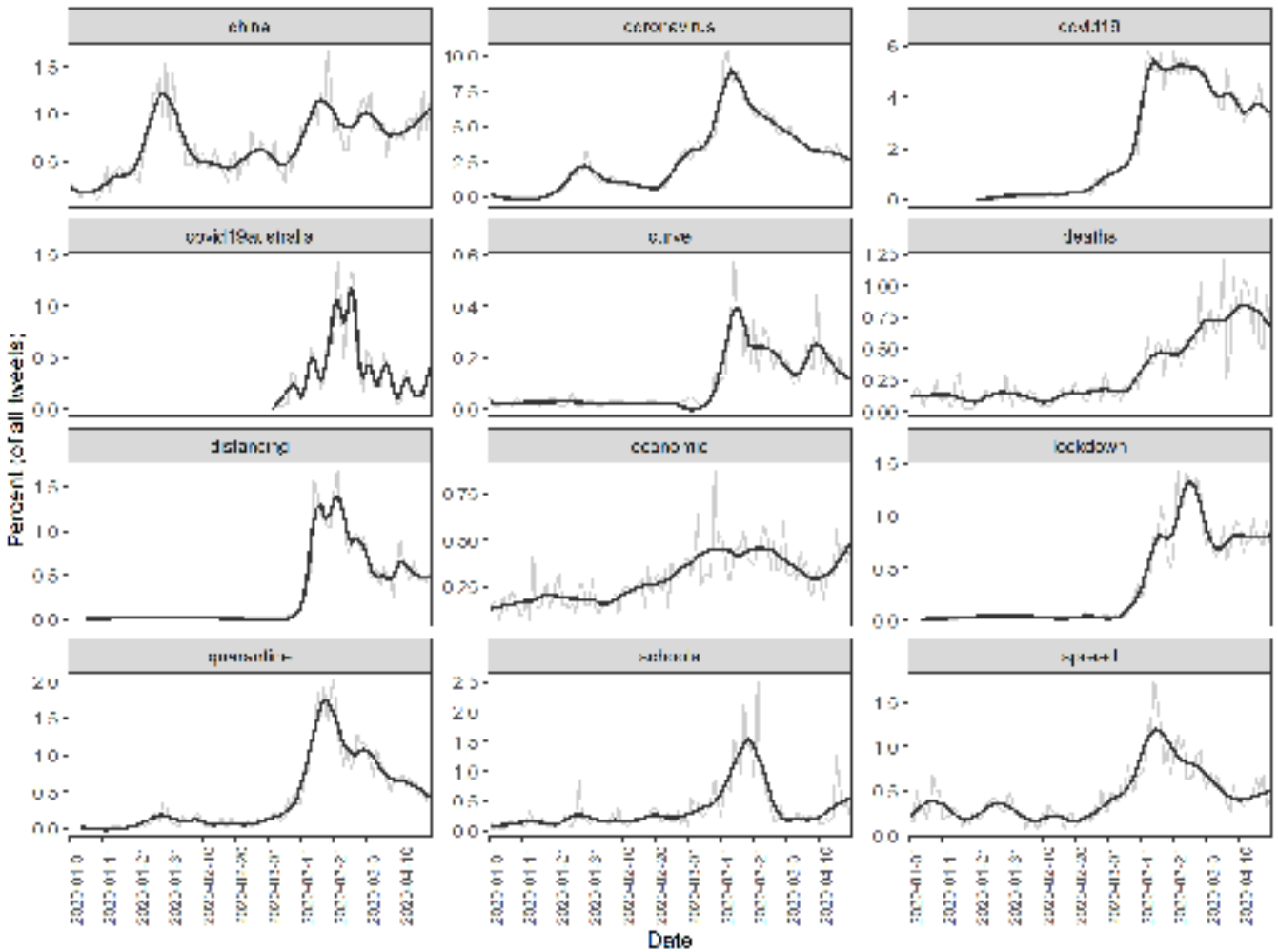

Figure 6: Linegraphs of selected COVID19-relates keyterms across time of COVID19-related keywords

Periodization

Identifying periods based on the frequencies of the keywords for each day

Problem

- How many periods are there?
- Are clusters/periods continuous?

Solution

- PAM clustering (partition around medoids) for 1:20 clusters
- Determining optimal partitioning solution (N clusters using Calinski-Harabasz scores (Łukasik et al. 2016))

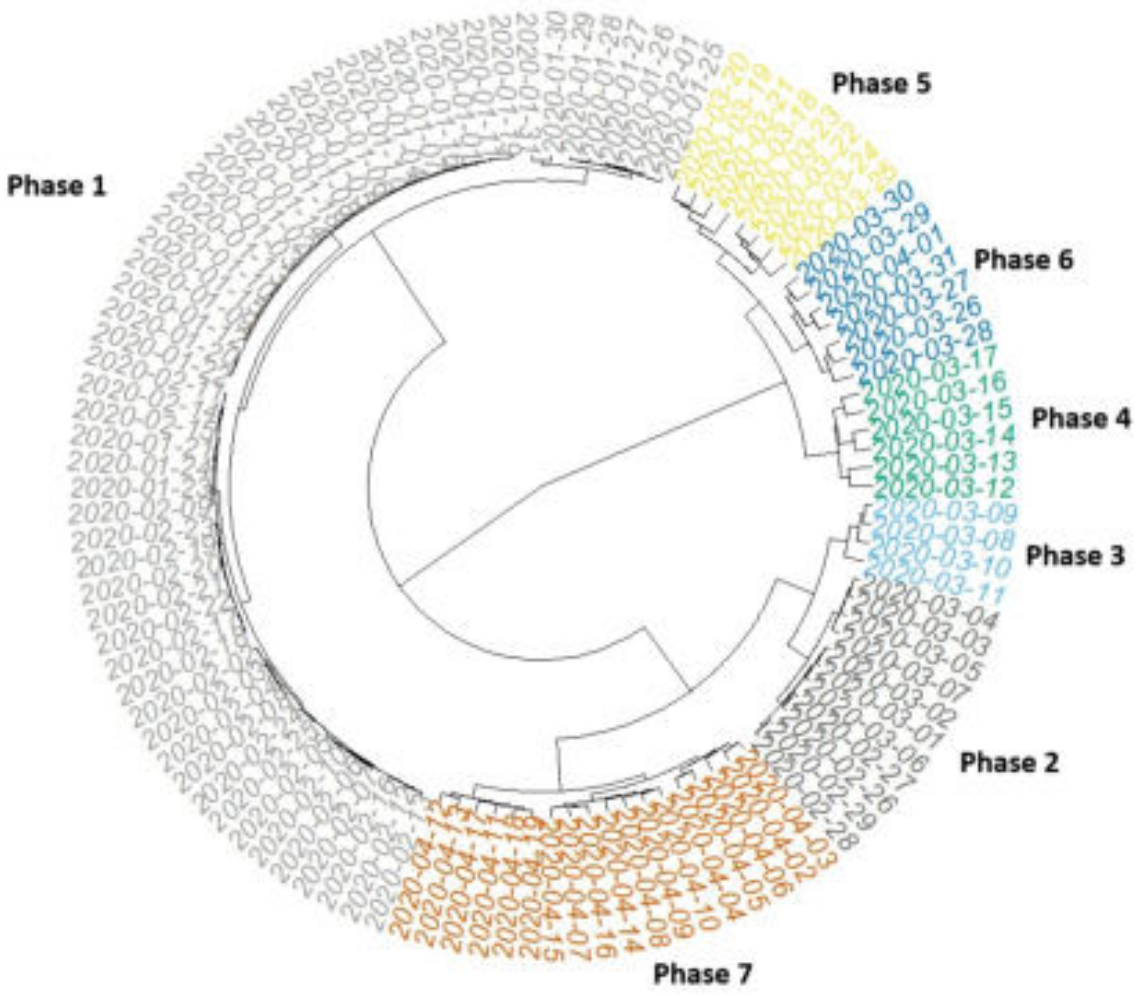

Figure 7: Results of the PAM clustering showing the data-driven periodization of the data

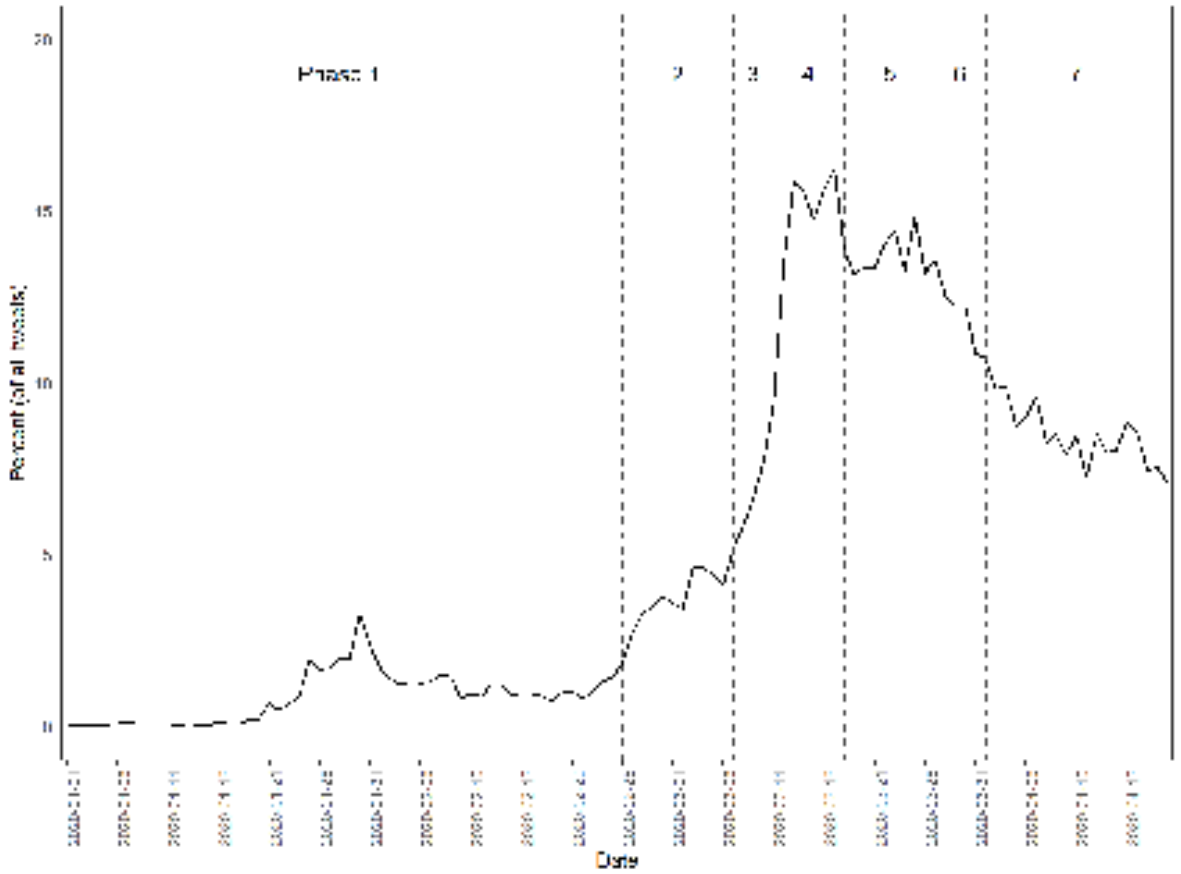

Figure 8: Percantages of COVID19-related tweets by period

Keywords per period

Identifying keywords that are significantly associated with one period

Solution

- Fisher's Exact tests (with Benjamini-Hochberg correction for multiple/repeated tests) (Field et al. 2012)
- Applied to all word types in the data $\rightarrow$ 49 terms that are significantly and positively correlated with COVID-19 discourse

	N Period	**N other Periods**
Element	a	b
Other elements	c	d

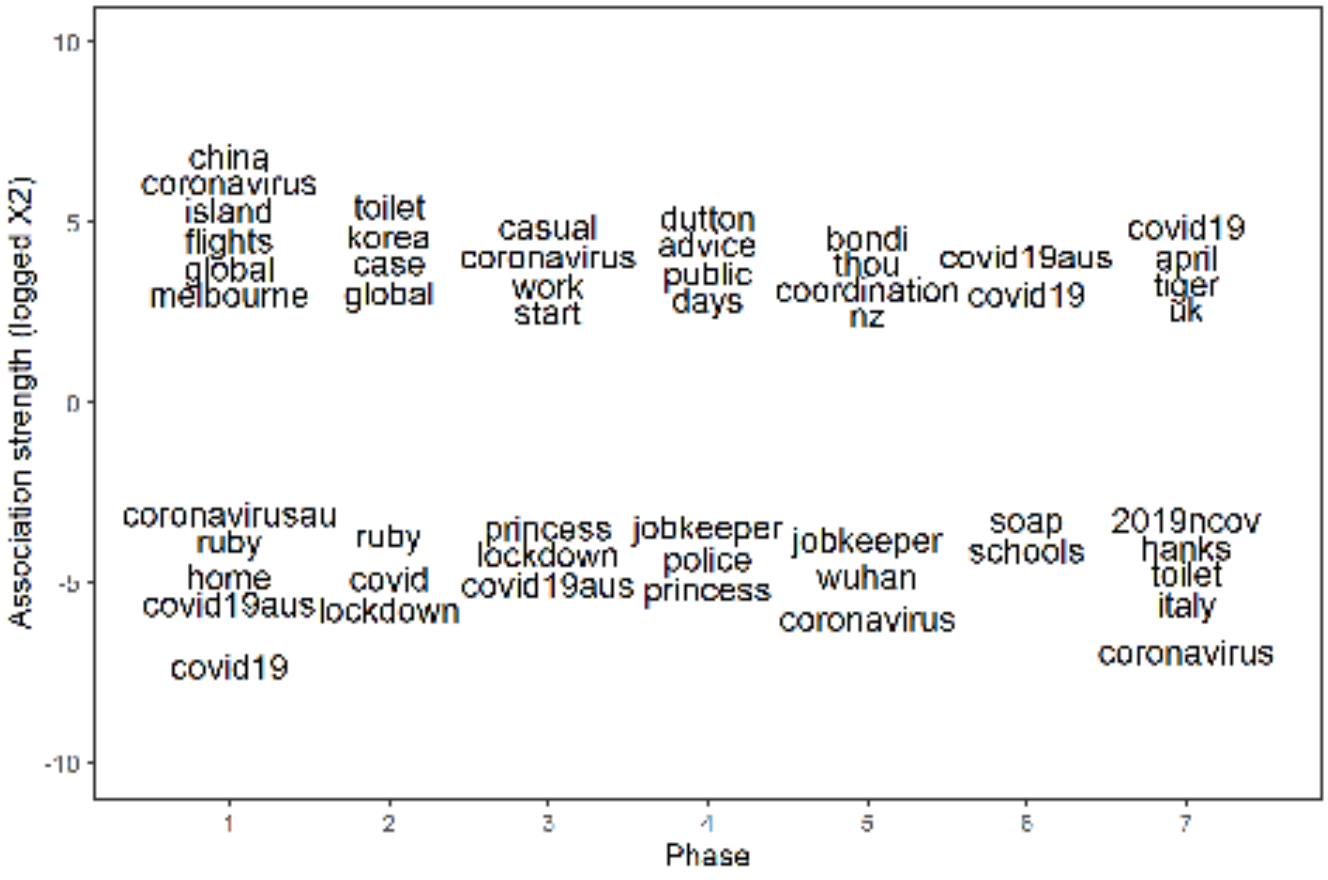

Figure 9: Words significantly over and underused across periods

Topics

Identifying topics in the COVID-19 discourse: LDA (Latent Dirichlet Allocation; Blei et al. (2003))

Problem

- Cohesiveness of topics?
- Optimal number of topics?

Solution

- 1:20 solutions (coherence scores; Cao et al. (2009), Deveaud et al. (2014))
- Fisher's Exact tests (with Benjamini-Hochberg correction for multiple/repeated tests) → as with key words

LDA topic model

Topic 1 MEDICAL	Topic 2 INTERNATIONAL	Topic 3 RESTRICTIONS/HOME	Topic 4 SPREAD	Topic 5 ECONOMY
sticking (.089)	trump (.074)	lockdown (.053)	positive (.067)	workers (.049)
tongue (.089)	cases (.073)	stay (.051)	tested (.066)	auspol (.048)
patients (.050)	china (.071)	home (.046)	cruise (.056)	support (.045)
erts (.039)	deaths (.070)	kids (.038)	princess (.054)	crisis (.043)
masks (.036)	chinese (.050)	love (.032)	ship (.050)	government (.041)
doctors (.035)	iran (.046)	toilet (.030)	nsw (.050)	economic (.040)
vaccine (.034)	death (.044)	shopping (.028)	ruby (.049)	package (.039)
covid19 (.032)	president (.043)	quarantine (.028)	passengers (.046)	stimulus (.038)
treatment (.029)	wuhan (.040)	day (.027)	minister (.039)	economy (.035)
care (.028)	italy (.040)	paper (.025)	sydney (.037)	pay (.035)

Ten most strongly associated keywords for each topic (values in round brackets represent ϕ (phi) to indicate association strength (all words were highly significant after Benjamini-Hochberg correction)

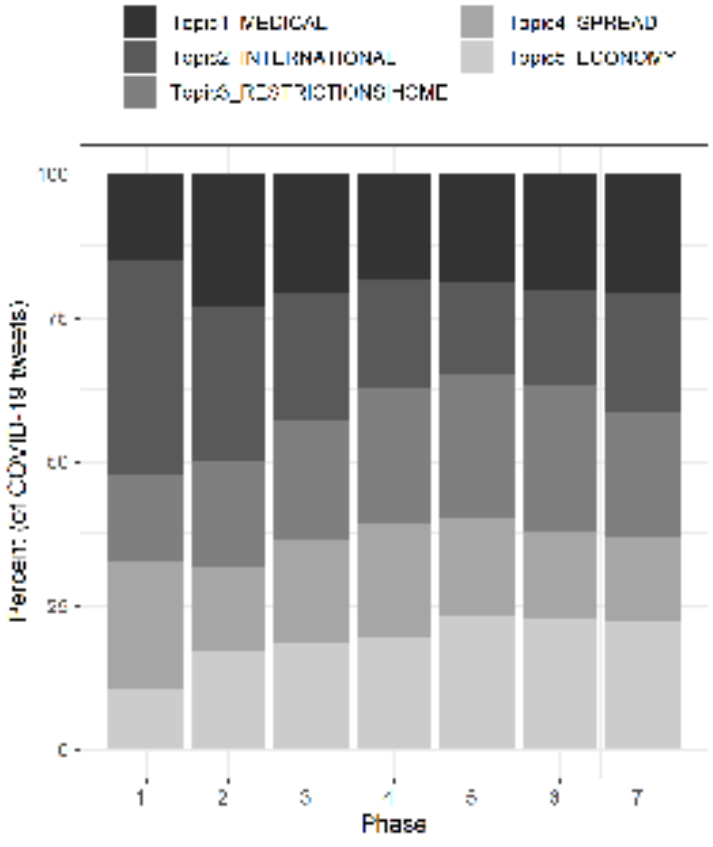

Figure 10: Distribution of topics across periods (bar plot)

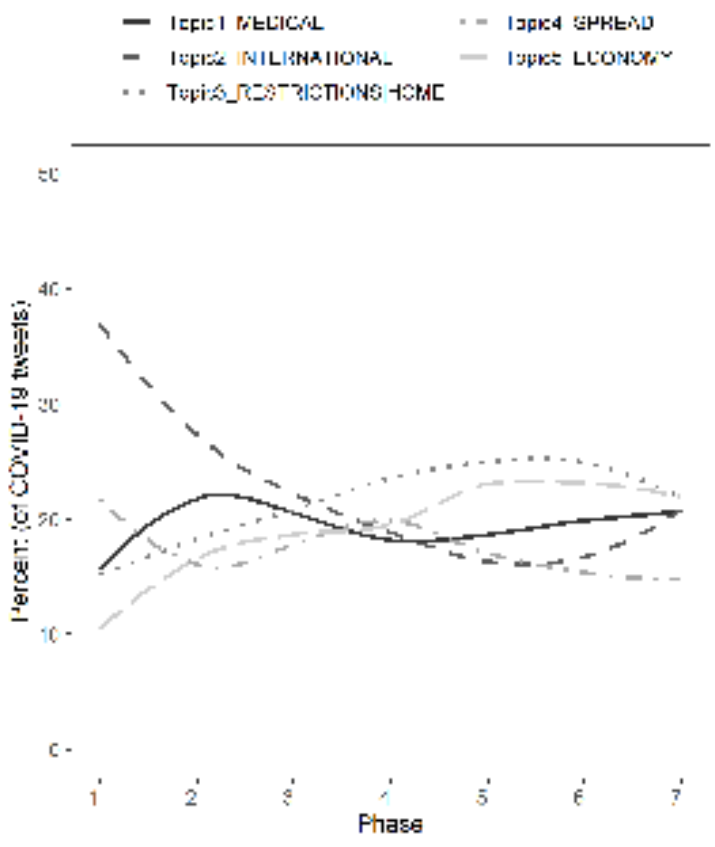

Figure 11: Distribution of topics across periods (loess smoothed)

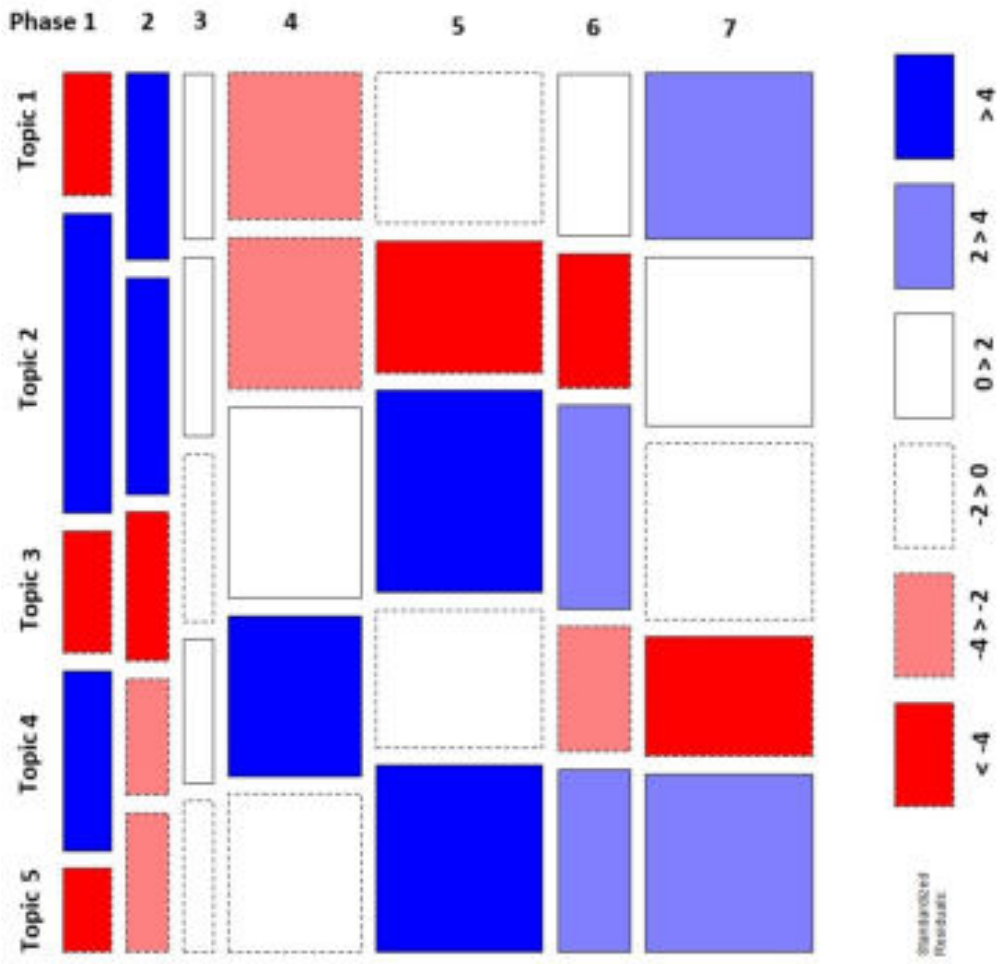

Figure 12: Mosaicplot of topics across periods

Sentiments

Identifying changes in polarity and emotionality in the COVID-19 discourse

Questions

- Does the discourse show shifts in polarity/stance/approval?
- Which topics are viewed positively/negatively?
- Which emotions are the topics associated with?

Problems

- Subjective ratings
- Ratings based on word lists

Solution

Sentiment analysis (Jockers 2015) based on *Word-Emotion Association Lexicon* (Mohammad and Turney 2013)

- 10,170 terms rated through crowd-sourced Amazon Mechanical Turk service (38,726 ratings, 2,216 raters)
- Associated with basic emotions (ANGER, ANTICIPATION, DISGUST, FEAR, JOY, SADNESS, SURPRISE, TRUST; see (Plutchik 1994))
- Each term rated 5 times (85%: 4+ identical ratings)
  - *dark* or *tragic*: SADNESS
  - happy or beautiful: JOY
  - cruel or outraged: ANGER
- ANGER|DISGUST|FEAR|SADNESS = negative
- ANTICIPATION|JOY|SURPRISE|TRUST = positive

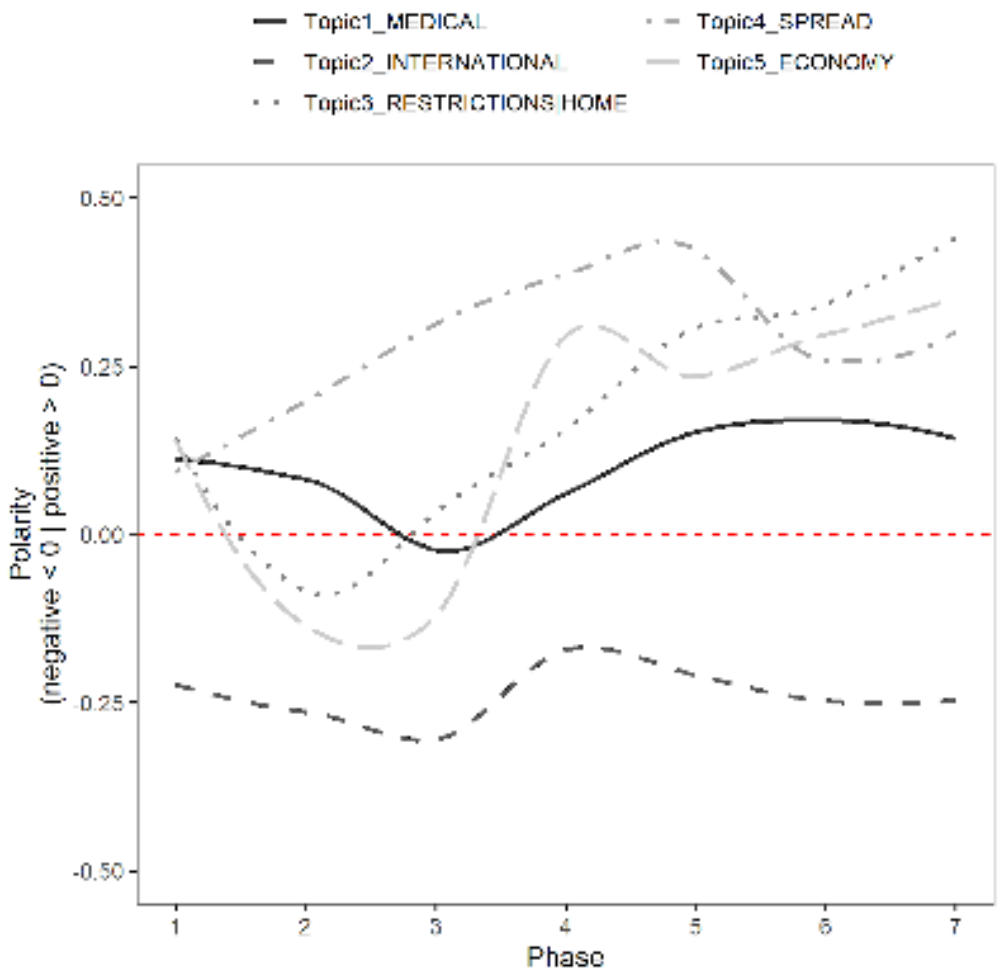

Figure 13: Polarity of topics across periods

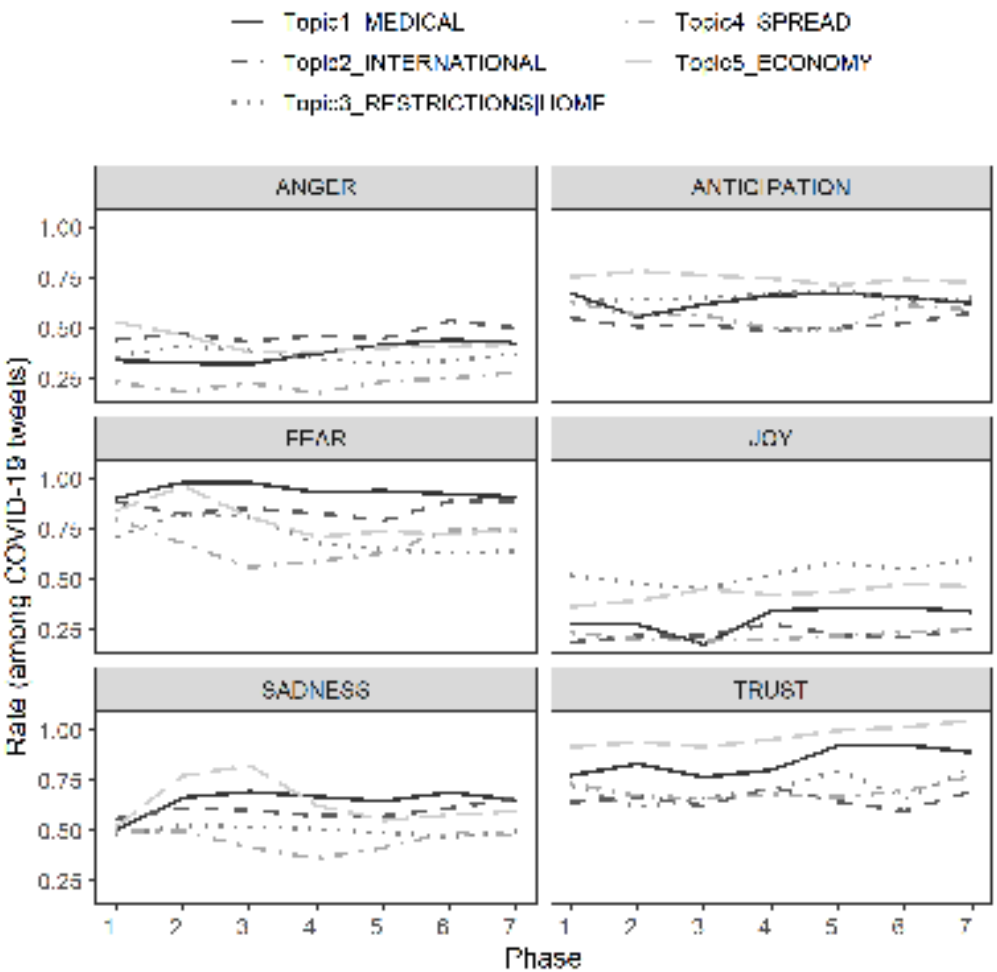

Figure 14: Emotionality of topics across periods

Outlook

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 entire data
- Extend period (beyond Apr. 15)
- Separate analyses for n-grams and then combine results (if combined before analysis, n-grams not significant)
- Apply same method to other phenomena (BLM, Bushfires) to get a better understanding of how public/Twitter discourse evolves over time

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.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(3), 993–1022.

Cao, J., X. Tian, L. Jintao, Z. Yongdong, and T. Sheng (2009). A density-based method for adaptive lda model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008 72*(7–9), 1775–1781.

Deveaud, R., r. SanJuan, and P. Bellot (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique 17*(1), 61–84.

Field, A., J. Miles, and Z. Field (2012). *Discovering statistics using R*. Sage.

Jockers, M. (2015). Package 'syuzhet'. access 2016/02/15.

Łukasik, S., P. A. Kowalski, M. Charytanowicz, and P. Kulczycki (2016). Clustering using flower pollination algorithm and calinski-harabasz index. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2724–2728. IEEE.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

Plutchik, R. (1994). *The psychology and biology of emotion.* Harper Collins College Publishers.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Text Mining the COVID-19 discourse in the Australian Twittersphere

Martin Schweinberger (UQ)
Michael Haugh (UQ)
Sam Hames (QUT)

slides available at

www.martinschweinberger.de

m.schweinberger@uq.edu.au

R code upon request after embargo