Text Mining the COVID-19 discourse in the Australian Twittersphere

Martin Schweinberger (UQ)
m.schweinberger@uq.edu.au

Michael Haugh (UQ)
michael.haugh@uq.edu.au

Sam Hames (QUT)
sam.hames@qut.edu.au

slides available at
www.martinschweinberger.de
R code upon request after embargo

Why analyze COVID-19 discourse on Twitter?

General

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

Previous research

- Predominantly by non-linguists
- 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)

Introduction

Data

Analysis

- Classification
- Keyword extraction
- Periodization
- Topic modeling
- Sentiment analysis

Discussion and Outlook

Starting point

Research Questions

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)?

Can the application of corpus linguistic methods unearth patters that were not identified by ML methods?

Hypotheses

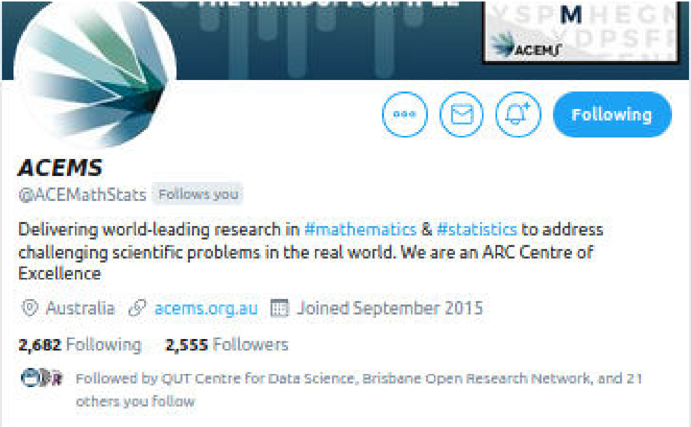
We assume that COVID-19 discourse developed in these phases:

Phase 1: China → Phase 2: Covid/Lockdown → Phase 3: Economy

Breaking down the discourse allow us to detect concepts that characterize periods that ML methods were not able to detect, because they conceptualized the COVID-19 discourse as an undifferentiated whole.

Data

What kind of Twitter data?



ACEMS

@ACEMathStats Follows you

Delivering world-leading research in #mathematics & #statistics to address challenging scientific problems in the real world. We are an ARC Centre of Excellence

Australia acems.org.au Joined September 2015

2,682 Following **2,555** Followers

Followed by QUT Centre for Data Science, Brisbane Open Research Network, and 21 others you follow

Figure 1: User information

How can you collect Twitter data?

Digital Observatory at QUT

- (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
→ Use training set to identify keywords and build a classifier

Data processing

Entire workflow in R (R Core Team 2020) for reproducibility

- 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

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) → length of vector: 16,087
- Test set: 1,250 tweets → 100 % prediction accuracy

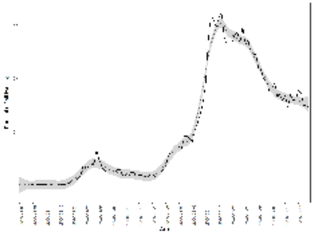


Figure 2: Percent of COVID19 tweets of all tweets per day

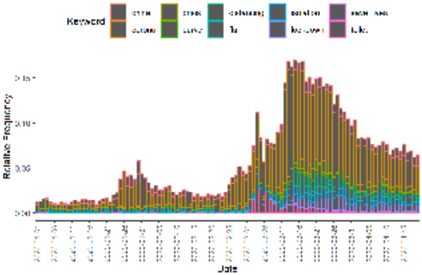


Figure 3: Percent of selected COVID19-related key terms per day

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 → 49 terms that are significantly and positively correlated with COVID-19 discourse

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

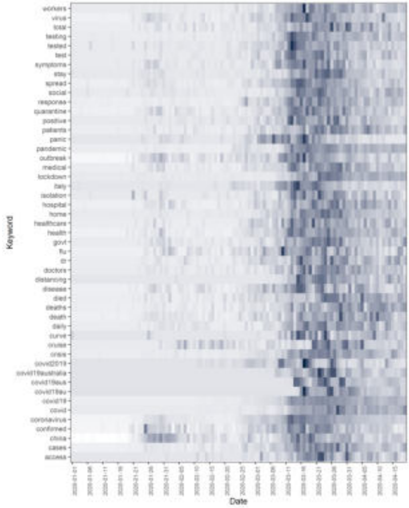


Figure 4: Heatmap of COVID19-related keywords

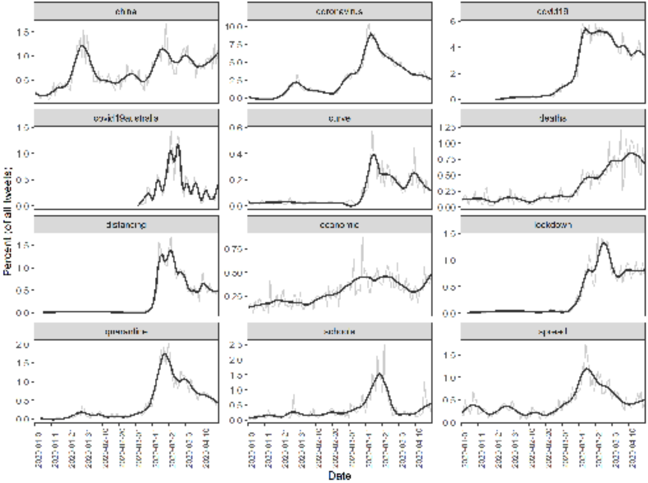


Figure 5: Linegraphs of selected COVID19-relates keyterms across time of COVID19-related keywords

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))

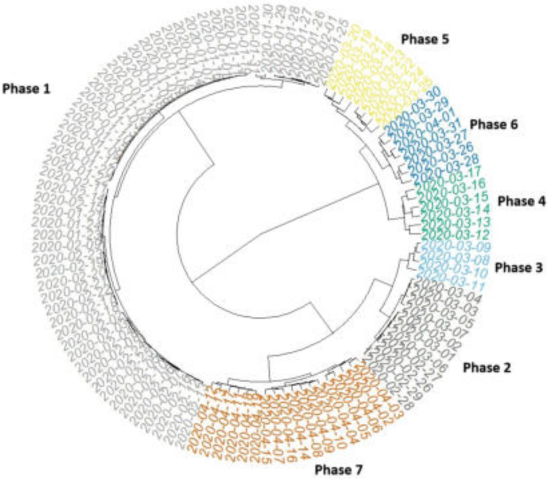


Figure 6: Results of the PAM clustering showing the data-driven periodization of the data

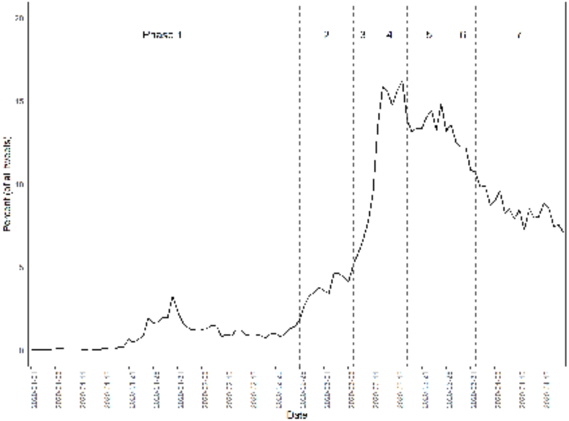


Figure 7: Percantages of COVID19-related tweets by period

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 → 49 terms that are significantly and positively correlated with COVID-19 discourse

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

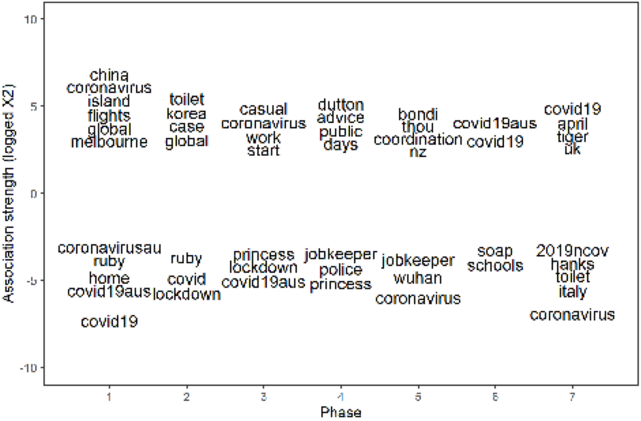


Figure 8: Words significantly over and underused across periods

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)

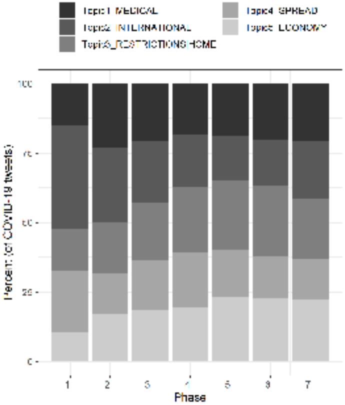


Figure 9: Distribution of topics across periods (bar plot)

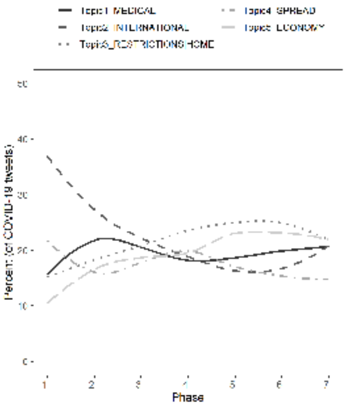


Figure 10: Distribution of topics across periods (loess smoothed)

Sentiments

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

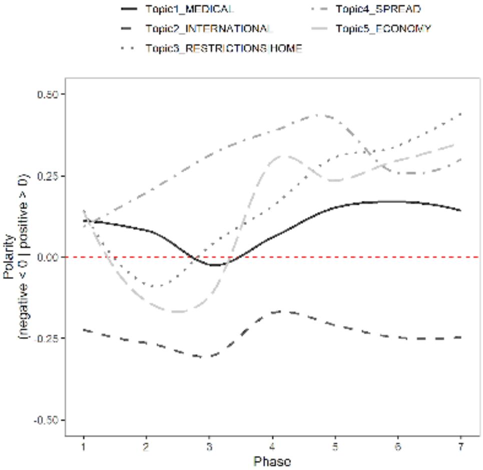


Figure 11: Polarity of topics across periods

Conclusion and Outlook

Conclusion

This analysis has shown how machine learning (ML) (classification, topic modeling, sentiment analysis) can be enhanced with corpus linguistic (CL) methods (key word extraction, cross-tabulation).

- Using ML-based approaches has only provided a rather superficial overview of the COVID-19 discourse
- Traditional CL methods, in isolation, provide detailed insights but they are limited to certain aspects (small data sets/few users)
- Breaking discourse down into phases and sub-discourses and combining ML and CL can help in providing a more detailed understanding of discourse around social phenomena and allows us to unearth patterns that would not be detectable otherwise

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)
m.schweinberger@uq.edu.au

Michael Haugh (UQ)
michael.haugh@uq.edu.au

Sam Hames (QUT)
sam.hames@qut.edu.au

slides available at
www.martinschweinberger.de
R code upon request after embargo