Talking about news articles related to the Covid Pandemic on Twitter

Global Discourse and Intertextuality

Sven Albrecht, Marina Ivanova

TU Chemnitz

30.08.2022

Funded by Deutsche Forschungsgemeinschaft German Research Foundation

Contact: sven.albrecht@phil.tu-chemnitz.de,marina.ivanova@phil.tu-chemnitz.de

About us

Sven Albrecht

- BA & MA in English and American Studies from TU Chemnitz
- worked as vocational school teacher in Germany
- worked as high school teacher in China
- currently working at TUC as part of the DFG funded CRC **Hybrid Societies** and the Erasmus+ project **TEFL-ePAL**

Marina Ivanova

- BA & MA in English and American Studies from TU Chemnitz
- PhD project measuring brain activity (EEG) to study how Slavic and German English learners perceive word stress
- integrate cues in a credible conversational pedagogical agent (CRC **Hybrid Societies** associate)
- Coordinator of DAAD project **CompConTrustOO**

Introduction

Vision

It would be really awesome if we could connect corpus data with social media data.

Research Questions:

- RQ1 How many distinct topics related to the Covid-19 pandemic can be found in the Twitter data?
- RQ2 How do people on Twitter refer to the news articles from the Coronavirus Corpus?
- RQ3 Which intertextual elements and fuctions are employed in Tweets referencing news articles from the Coronavirus Corpus?

Intertextuality

Definition

The act of texts referencing other texts.

- literary origins of the term (Kristeva, Bakhtin) and incorporated in CDA (Fairclough, 1992)
- narrow sense: textual overlap of the user's tweet and the headline
- broad sense: the expression of personal comments referencing the topic
- recent accounts on Corona responses in the media and Twitter indicate dialogue and intertextuality (Dong, Buckingham, & Wu, 2021; Kurten & Beullens, 2021; Schweinberger, Haugh, & Hames, 2021; Tsao et al., 2021)

Did somebody say "Big Data"?

Coronavirus Corpus (Davies, 2021)

- News articles scraped from online newspapers and magazines in 20 different English-speaking countries
- ∼1 Million articles at the time
- ∼869 Million words

Covid-19 Twitter Data (Banda et al., 2022)

- Tweets containing "coronavirus", "2019nCoV", "corona virus", later "COVD19", "CoronavirusPandemic", "COVID-19", "2019nCoV", "CoronaOutbreak", "coronavirus", "WuhanVirus"
- ∼350 Million tweets (excluding retweets)
- ∼1.6 Million tweets containing URLs of articles form Coronavirus Corpus

Handling Big Data

Challenges

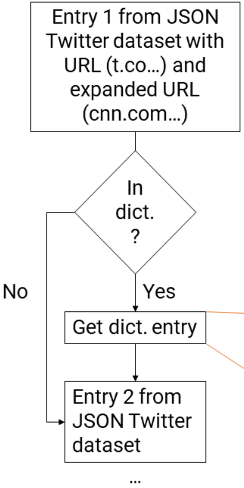
- hydrating Twitter data takes weeks
- size of Twitter Corpus: 1.2TB uncompressed JSON data
- runtime of analysis scripts matching URLs in the two data sets

Solutions

- running hydration tool on a server in a tmux session
- on-the-fly gzip compression of incoming data
- pickles, dictionary look-ups instead of list comprehension, parallelization

Code

All code used in the analysis is availble in our Gitlab repository: `https://mytuc.org/vkjk`



Dictionary

Keys (URLs from CoronaCorp)	Values (headlines from CoronaCorp)
'https://www.sfgate.com/news/article/Baker-unveils-171M-plan-aimed-at-aiding-tenants-15640748.php':	'Baker unveils $171M plan aimed at aiding tenants, landlords'
‘https://www.cnn.com/2020/05/10/politics/mike-pence-self-isolate-coronavirus/index.html’:	Pence will not self-quarantine and plans to be at the White House Monday

Figure 1: Illustration of the dictionary lookup procedure

Topic Modeling

Theoretical Assumptions

- Saussurean stance that meanings are relational (Mohr & Bogdanov, 2013)
- text as *bag-of-words*, disregarding all other complexities such as syntax, semantics, structure, word order (DiMaggio, Nag, & Blei, 2013; Mohr & Bogdanov, 2013)

Various unsupervised probabilistic machine learning algorithms available:

- latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003)
- latent semantic indexing (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 1991)
- probabilistic latent semantic indexing (Hofmann, 1999)
- modified LDA (Blei & Lafferty, 2006; Chang & Blei, 2010; Griffiths, Steyvers, Blei, & Tenenbaum, 2004; Wallach, 2006)

Latent Dirichlet Allocation

Blei, Carin, and Dunson (2010)

"[LDA is] a hierarchical probabilistic model used to decompose a collection of documents into salient topics, where a 'topic' for LDA is a probability distribution over a vocabulary"

- each document is a distribution of topics
- every topic is a distribution of words
- only documents and words are observed variables
- topics are latent variables

LDA as reverse engineering of (imaginary) generative process of documents:

- fixed number of topics in the corpus
- each document exhibits these topics to a varying degree (Blei, 2012)

Current Study

Preprocessing

- removal of stop words (incl. coronavirus & covid*)
- removal of words < 3 characters
- word frequency > 50% and < 10 removed
- words lemmatized and stemmed

Analysis

- LDA implementation of Gensim (Rehurek & Sojka, 2010) in Python
- hyperparameter optimization (α, η, κ, τ_0, see Hoffman, Bach, and Blei (2010))

Estimating the number of topics

- coherence (Blei et al., 2003; Mimno, Wallach, Talley, Leenders, & McCallum, 2011)
- perplexity (lower score = better performance, see Wallach, Murray, Salakhutdinov, and Mimno (2009))

Intertextuality Annotation

3057 tweets were annotated:

- Overlap (manual): yes / no / paraphrase
- Comment (manual): yes / no
- Mentions and Hashtags (automatic)
- Personal account (semi-automatic)

URL	headline	tweet	overlap	comment	mentions	hashtags	verified	tw_ID	username	user_display
https://www.cnn.com/world/live-news/coronavirus-outbreak-03-11-20-intl-hnk/index	March 11 coronavirus news - CNN	On behalf of @KenyaMedics_KMA I send our condolences to Italian medical fraternity on death of Roberto Stella, president of the Medical Guild of Varese, died Tuesday night in Como of #Coronaitaly @mdjkitulu @LukoyeAtwoli @SupaTunje @JKARAMANA @lizzgitau https://t.co/W8oLcEXlwz	no	TRUE	TRUE	TRUE	personal account	1.24E+18	simonkigondu	The Kenyan Gyne #RejectHealthLawsAmendmentBill2021
https://www.cnn.com/interactive/2020/health/coronavirus-us-deaths-milestones/	Understanding the massive scale of coronavirus in the US	Understanding the massive scale of coronavirus in the US via this truly beautiful and mobile friendly interactive: https://t.co/t1DBY3ETxc	yes	FALSE	FALSE	FALSE	journalist	1.27E+18	EricaAlyssa	Erica A. Hernandez

Figure 2: Sample of the annotation dataset

Number of Topics

Hyperparameters: $\alpha = 0.51$, $\eta = 0.51$, $\kappa = 0.5$, $\tau_0 = 1$, $chunksize = 16384$

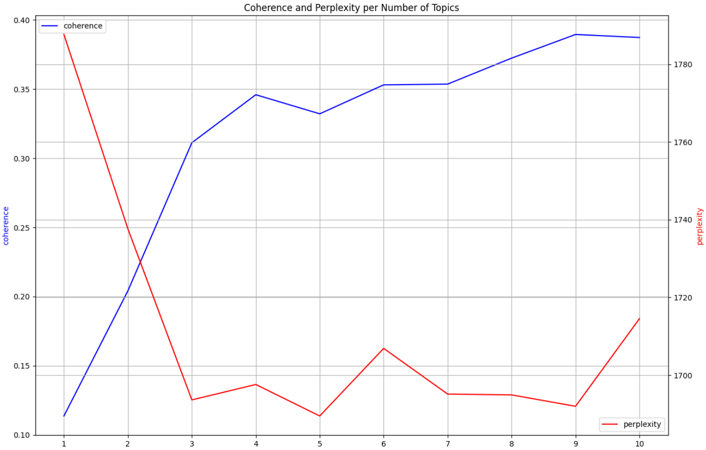


Figure 3: Example of Perplexity and Coherence Scores for 1-10 Topics

Intertopic Distance Maps

Figure 4: Visualization for four topics

Figure 5: Visualization for nine topics

Top-10 Most Salient Terms – Four Topics

Topic 1 (26.9% of tokens)	Topic 2 (26.3% of tokens)	Topic 3 (24.1% of tokens)	Topic 4 (22.7% of tokens)
test	vaccin	trump	cas
peopl	pandem	mask	stat
lik	heal	say	death
posit	spread	am	tim
vir	sci	wear	report
die	effect	respons	infect
study	pfiz	presid	week
think	publ	hous	increas
year	work	fauc	york
read	world	realdonaldtrump	new

Table 1: Ten most frequent words for four topics

Top-10 Most Salient Terms – Nine Topics

Topic 1 (12.5%)	Topic 2 (12.1%)	Topic 3 (11.4%)	Topic 4 (11.4%)	Topic 5 (11.2%)	Topic 6 (10.6%)	Topic 7 (10.5%)	Topic 8 (10.3%)	Topic 9 (10.1%)
work	peopl	trump	vaccin	cas	mask	die	stat	year
heal	lik	pandem	test	tim	wear	off	govern	sci
risk	know	am	posit	death	fac	travel	nat	week
nee	go	read	spread	new	fauc	vary	hous	warn
med	get	artic	vir	report	school	chief	reliev	mil
hom	think	com	hospit	infect	check	march	whit	country
car	tel	repons	research	dat	plan	book	emerg	liv
help	dont	presid	study	increas	busy	serv	cal	chin
publ	good	lead	effect	york	clos	break	elect	world
long	look	realdonaldtrump	pfiz	numb	op	county	repub	dea

Table 2: Ten most frequent words for nine topics

Comparison with previous studies

Topic models of 373,908 tweets (25.02-30.03.2020) from Belgium in English

1	2	3	4	5	6
coronavirus	coronavirus	coronavirus	coronavirus	coronavirus	coronavirus
covid19	covid19	covid19	covid19	covid19	covid19
crisi	corona	time	corona	crisi	corona
countri	peopl	take	impact	support	outbreak
pandem	need	peopl	pandem	test	peopl
peopl	support	spread	need	work	need
measur	work	help	today	belgium	spread
fight	belgium	european	help	show	fight
help	health	call	social	good	test
member	itali	case	close	itali	european

Table 3: Topic models of Kurten and Beullens (2021, 120)

Comparison with previous studies

Topic models of 769,165 tweets (01.01-20.04.2019 and 01.01-20.04.2020) from Australia (Schweinberger et al., 2021)

- medical
- international
- restrictions|home
- spread
- economy

Intertextuality: Quantitative

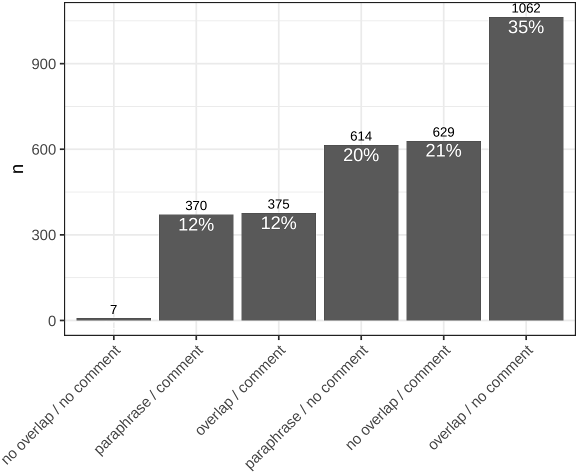


Figure 6: Distribution of overlap types and comments

Intertextuality: Quantitative

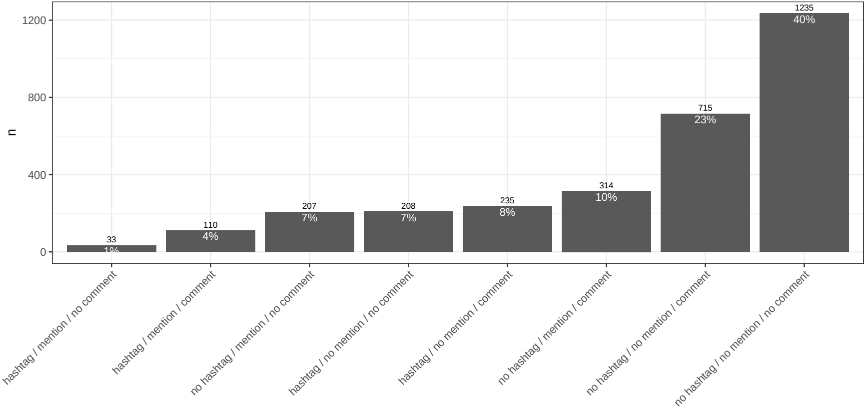


Figure 7: Distribution of hashtags, mentions and overlap types

Intertextuality: Quantitative

96% personal accounts, 4% verified accounts:

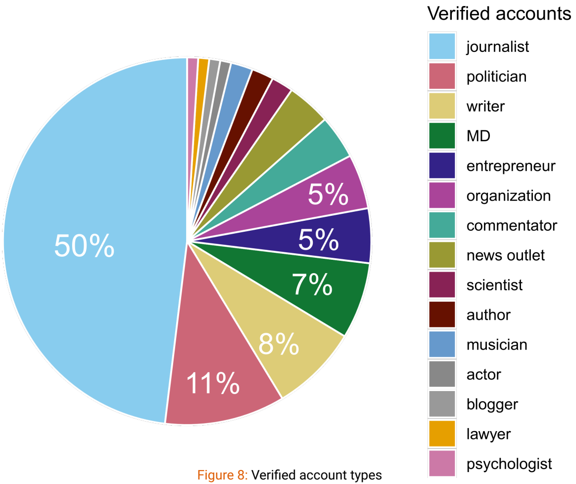


Figure 8: Verified account types

Verified account	Overlap	Comment	N	Percent
journalist	overlap	no comment	21	16%
journalist	paraphrase	comment	16	12%
journalist	paraphrase	no comment	16	12%
journalist	no overlap	comment	10	8%
politician	no overlap	comment	6	5%
writer	no overlap	comment	5	4%
MD	paraphrase	comment	4	3%
politician	paraphrase	comment	4	3%
...				

Table 4: Verified account intertextual behavior

Intertextuality: Qualitative

Comment types

- **advertising**

H: "Coronavirus: Fact vs Fiction - Podcast on CNN Audio"

T: "Good, solid, basic info separating fact from fiction on novel coronavirus by @drsanjaygupta @cnn. https://t.co/G6nBzRlFNR

- **criticising**

H: "Pence will not self-quarantine and plans to be at the White House Monday"

T: "Wrong decision. Bad example. Crap leadership. Pence will not self-quarantine and plans to be at the White House Monday https://t.co/AmVWWZhmky"

- **supporting**

H: "Dutch leader did not visit dying mother for weeks to comply with coronavirus lockdown"

T: "A real leader leads by example, especially during a crisis: Dutch leader did not visit dying mother for weeks to comply with coronavirus lockdown https://t.co/Sw0qKDWdYS"

Intertextuality: Qualitative

Intertextual elements

- **deictics**

H: "At Least 128,000 People in the U.S. Have Received the Covid-19 Vaccine"

T: "@SenFeinstein @KamalaHarris Im a loyal Dem Why is CA so low on this list (it is sorted by %)? #COVID19 #vaccines https://t.co/PM8vuQwCdH"

- **reported speech**

H: "Pfizer and BioNTech say their coronavirus vaccine was 95%"

T: "@natvallade @matthewdmarsden @wildpinkrabbit @joerogan When the Pfizer and Moderna vaccines were introduced at the end of 2000, prior to mass rollout, the claim from the vaccine makers, accepted by CDC, was that they were 95% effective in preventing infection and transmission: https://t.co/TmDn0eTGYP"

- **changed headlines**

H: "Stop touching your face all the time to avoid spread of the coronavirus. It's easier said than done"

T: "One big coronavirus challenge is how to stop touching your face https://t.co/A46uiZDtBB".

- **dynamic headlines**

H: "Coronavirus Live Updates: Trump Aides Target Fauci"

T: "Great idea! Live Coronavirus Updates: 17 States Sue Trump Administration - The New York Times https://t.co/kGT9q7394l"

Limitations

- handling "big data" (storage space, run time of analysis, limitations of available tools)
- Twitter API limits (data collection via streaming API, hydration rate limits)
- topic modeling not deterministic (models available at: `https://mytuc.org/btkk`)
- methodological considerations (bag-of-words approach, interpretation of results by researcher, manual analysis of intertextuality)

Conclusion

- Combining corpus data with social media data is possible and feasible for linguistic research
- Topic Modeling suggested four or nine distinct topics in the Twitter data revolving around testing, vaccines, political figures, and infection statistics
- Intertextuality patterns on Twitter reflect fast discourse: users mostly either retweet the headline without change or completely omit it and include a comment
- Most tweets lack hashtags and mentions – isolated discourse on critical topics
- Intertextual devices are used to advertise, criticise and support news in social media

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., . . . Chowell, G. (2022, April). *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.* Zenodo. doi: 10.5281/ZENODO.6481639
Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1), 8–11.
Blei, D. M., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, *27*(6), 55–65.
Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning (ICML'06)*, 113–120.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.
Chang, J., & Blei, D. M. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, *4*(1), 124–150.
Davies, M. (2021, November). The Coronavirus Corpus: Design, construction, and use. *International Journal of Corpus Linguistics*, *26*(4), 583–598. doi: 10.1075/ijcl.21044.dav
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.
DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570–606.
Dong, J., Buckingham, L., & Wu, H. (2021). A discourse dynamics exploration of attitudinal responses towards covid-19 in academia and media. *International Journal of Corpus Linguistics*, *26*(4), 532–556.
Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior research methods, instruments, & computers*, *23*(2), 229–236.
Fairclough, N. (1992). Intertextuality in critical discourse analysis. *Linguistics and Education*, *4*(3-4), 269–293.
Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. *Advances in neural information processing systems*, *17*.
Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, *23*.
Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50–57).
Kurten, S., & Beullens, K. (2021). #coronavirus: Monitoring the Belgian twitter discourse on the Severe Acute Respiratory Syndrome Coronavirus 2 pandemic. *Cyberpsychology, Behavior, and Social Networking*, *24*(2), 117–122.
Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).
Mohr, J. W., & Bogdanov, P. (2013). *Introduction—Topic models: What they are and why they matter* (Vol. 41) (No. 6). Elsevier.
Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.*
Schweinberger, M., Haugh, M., & Hames, S. (2021). Analysing discourse around COVID-19 in the Australian twittersphere: A real-time corpus-based analysis. *Big Data & Society*, *8*(1), 20539517211021437.
Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, *3*(3), e175–e194.
Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977–984).
Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).

Talking about news articles related to the Covid Pandemic on Twitter

Global Discourse and Intertextuality

Sven Albrecht, Marina Ivanova

TU Chemnitz

30.08.2022

HYBRID SOCIETIES

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Contact: `sven.albrecht@phil.tu-chemnitz.de,marina.ivanova@phil.tu-chemnitz.de`