“Mining Twitter to Detect Hotspots in Psychology”

Bittermann, A., Batzdorfer, V., Müller, S. M., & Steinmetz, H.

(ZPID – Leibniz Institute for Psychology)

PsychArchives-ESM 1 (Methods)

In this supplementary material, additional methodological details are described.

Sampling Process

As a first step, we used a search engine for universities provided by the German newspaper “DIE ZEIT” (https://studiengaenge.zeit.de/) in order to collect all public and equivalent universities with a psychology department in Germany, Austria, and the German-speaking part of Switzerland. Based on these results, we gathered psychology departments and associated professors from respective university websites. Thereby we included substitute, visiting, extraordinary, assistant, and honorary professors and excluded former and affiliate members, associated scientists and professors, non-active emeritus/retired professors, and visiting scholars. Similarly, we looked for the research institutes on GERiT – German Research Institutions (https://www.gerit.org/) of the German research foundation DFG. We filtered for psychological research institutes and looked for associated professors. Regarding DGPs sections, we referred to the official website (https://www.dgps.de/index.php?id=48).

As a second step, we identified corresponding Twitter accounts for each entry on our list and for all sections of the German Psychological Society. For professors, we used the search string: „[name] AND twitter psychology [city]“. In order to make sure that we did not miss any accounts, we always checked with the search string „[university] AND twitter psychology“. This way, we were able to find persons using a pseudonym as their screenname, but used their real name for display name or gave revealing information in their profile description.

Handling of Twitter Data

In terms of ethical considerations and in line with the General Data Protection Regulation (GDPR) we particularly ensured anonymity, data-sparsity and data confidentiality of the obtained Twitter data. Relating to anonymity, the risk of identification of individual users is reduced as much as possible as no person-related content (e.g. individual tweets) or user information is published within the study. Another central aspect to anonymity is to separate account-based personal information from the data set so that two separate data sheets which are separately stored result, one with account handles and another one containing the respective personal identification IDs. Regarding the sparsity aspect, only data features that are essential to our research endeavor are processed, as well as stem from non-vulnerable accounts and contain non-sensitive content. Regarding the confidentiality aspect, data are kept on a backed-up, virus protected server which is accessed by a password-secured institute laptop. In that way unauthorized access of Twitter data is prevented. Further in line with confidentiality is that access to Twitter data is logged; i.e. the entry, modification and deletion of data is recorded. Additionally, access control is based on an authorization concept in which exclusively authorized staff members of the project have access to the Twitter data.

Selection of Relevant Bigrams

To separate relevant tweet content from the noise of social media communication, tweets were annotated using two lists: (1) All hashtags in the corpus (which themselves can be regarded as annotations made by the user), and (2) the most frequent relevant bigrams. Specifically, these lists served as whitelists for term inclusion: Instead of defining corpus-specific stopwords, we determined terms that are *not* dropped from the corpus and thus kept for subsequent topic modeling. In contrast to unigrams or trigrams, manual inspection favored the use of bigrams in addition to the hashtag list. Bigram relevance was determined by consulting the APA thesaurus (Tuleya, 2007). Two authors discussed which terms are relevant to psychological research according to this thesaurus. For illustration, the following list shows the 100 most common bigrams and their frequencies in tweets with selected bigrams printed in bold:

new_paper	894	et_al	586	looking_forward	552
please_rt	461	finden_#ff	443	special_issue	438
open_science	374	#stellenangebote_u.a	366	akt_#stellenangebote	327
join_us	300	**open_access**	299	**mental_health**	298
new_study	249	new_preprint	236	prof_dr	214
social_media	211	now_available	200	phd_student	183
blog_post	161	**registered_reports**	160	interesting_read	157
big_data	153	new_article	151	postdoc_position	151
call_papers	150	now_open	144	**cognitive_neuroscience**	140
summer_school	138	check_new	134	phd_position	133
new_work	130	**social_psychology**	127	peer_review	122
aktuelle_#stellenangebote	120	vielen_dank	120	please_share	120
mobile_brain	119	**decision_making**	118	herzlichen_glückwunsch	117
individual_differences	116	psychological_science	115	looks_like	115
phd_students	114	**brain_body**	112	mehr_#psyndex	111
great_work	110	new_research	109	mehr_infos	107
clinical_psychology	107	**body_imaging**	106	ab_uhr	106
right_now	100	get_touch	100	finden_#psyndex	100
new_review	100	next_week	100	feel_free	99
climate_change	99	**machine_learning**	98	#jobs_#psychology	98
working_memory	97	good_news	96	new_blog	96
human_brain	95	just_published	95	help_us	94
phd_positions	93	#openscience_movement	93	become_part	93
come_join	92	now_online	92	part_#openscience	92
open_data	91	via_@spiegelonline	91	**replication_crisis**	90
early_career	90	come_work	90	movement_current	90
weitere_infos	89	**effect_size**	89	openings_#jobs	89
years_ago	88	well_done	88	can_help	87

spread_word 87	work_us 87	**emotion_regulation** 85
free_access 84	first_time 83	thanks_tweet 82
social_distancing 82	current_zpid 82	zpid_openings 82
new_post 81	please_retweet 81	can_found 81
max_planck 81	great_news 80	**effect_sizes** 80
abstract_submission 80		

Topic Modeling

For identifying topics within the corpus of annotated tweets, we used a topic modeling variant specifically designed for short texts: the Biterm Topic Model (BTM; Yan, Guo, Lan, & Cheng, 2013). Unlike the popular Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), which models word occurrences in a *document*, BTM models biterm (i.e., pair of words) occurrences in a *corpus*. This solves the problem of sparse word co-occurrence patterns in tweets and thus yields better results. In direct comparison with other topic model variants, BTM results proved to be superior (Jonsson & Stolee, 2015; Yan et al., 2013). In addition, we manually compared the results of LDA vs. BTM on our corpus (for a fixed number of 30 topics, alpha = 0.01, k = 1/*k*) with the result of BTM topics being more semantically coherent (coherence_{LDA} = -163.3297; coherence_{BTM} = -150.6887; based on top 10 terms).

Although topic modeling is an unsupervised machine learning technique, some parameters have to be set prior to analysis, with the number of topics to be found being the most challenging. To determine the optimal number of topics in our tweet corpus, we followed the best-practice recommendations by Maier et al. (2018) and investigated several candidate models using different numbers of topics (k = 25 - 50), different values for hyperparameter alpha (0.001 and 0.01), and different random seeds for Gibbs initialization. Hyperparameter delta was fixed to 1/k. The range of k and alpha, respectively, was determined in pretests on sample data. For each k, the model with the highest mean of semantic coherence (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) and term exclusivity (Roberts et al., 2014) was selected and inspected manually regarding topic interpretability and semantic validity (Maier et al., 2018). From the final model with k = 46 topics, six topics had to be excluded as they were unstable across multiple inference runs (i.e., no topic reliability sensu Maier et al., 2018). Additional 19 topics were excluded as they were related to specific departments or institutions, subject recruitment, job offers, conference locations, or uninterpretable. Thus, 21 topics were included in the final analysis.

Topic Labels and Relevant Terms

As BTM topics are based on biterm occurences in the whole corpus, the most probable terms of a topic do not necessarily need to be relevant to the topic's key content. For example, a topic referring to the COVID-19 pandemic (see Topic 4 in Table 1) can comprise terms like "germany" and "study", which themselves are very unspecific and meaningless without terms like “covid-19” or “corona”. Thus, for inspecting temporal trends of the topics, only the most semantically meaningful terms according to the topic labels were used for selecting tweets and publications, respectively. For determining these relevant terms, we first created topic labels by a joint examination of most probable topic terms and most representative tweets for each topic (the tweets with the highest probability for each topic). Next, we discussed which topic terms best reflect the topic labels. For selecting tweets and PSYNDEX publications, these “relevant terms” were combined using boolean operators similar to literature search in databases. For instance, tweets addressing a topic on “Workplace Aging” (see Table 1), should not contain the term “aging” alone. Thus, the respective search string was: (“work” OR “job” OR “workplace”) AND (“aging” OR “retirement”). For all search strings, see the analysis code in PsychArchives-ESM 2. The topic labels and relevant topic terms were also used for investigating whether topics identified in tweets were also discussed at conferences.

Forecasting

In this study, we employed ARIMA (autoregressive integrated moving average) models, as they present for most modeling approaches and forecasting goals with time series the most flexible yet powerful option to account for trends, seasonality, and autocorrelation (see Jebb et al., 2015). Consequently, they are applied in a variety of scientific fields, such as in economics (e.g., forecasting prices and economic development), political science (e.g., forecasting votes), epidemiology (e.g., forecasting infection rates, health, and mortality), or climate research (e.g., forecasting climate). In psychology, the approach is slowly entering the field, as more researchers are able to collect intensive longitudinal data (Jebb et al., 2015).

That having said, we did not ignore that there are other options as well. In particular, we considered to model a Gompertz growth model (Franses, 1994) as well as exponential smoothing (ETS) models as further options (see Hyndman & Athanasopoulos, 2018). Of these, we dismissed the Gompertz model after some exploration due to the observable mismatch with the series. The Gompertz function has a S-shape and is a monotonic function (i.e., it increases albeit with different rates across time). Hence, it is an optimal approach for growth processes (e.g., infection rates). The series investigated in our paper, in contrast, showed varying numbers of publications with ups and downs across time. We did, however, closely inspect the difference performance of ARIMA versus ETS models by formally comparing the differences in data fit for all our series. In this regard, we followed recommendations by Hyndman and Athanasopoulos (2018) to rely on the Akaike Information criterion corrected for small sample sizes (AICc) to select the most valid model. We realize that an out-of-sample forecasting cross validation approach would have been more optimal, however, the rather short series in our paper prevented us from doing so as the number of information would have led to unreliable and inaccurate results. In this regard, Hyndman and Athanasopoulos noted that

> *"Ideally, we would test if our chosen model performs well out-of-sample compared to some simpler approaches. However, with short series, there is not enough data to allow some observations to be withheld for testing purposes, and even time series cross validation can be difficult to apply. The AICc is particularly useful here, because it is a proxy for the one-step forecast out-of-sample MSE. Choosing the model with the minimum AICc value allows both the number of parameters and the amount of noise to be taken into account"* (Hyndman and Athanasopoulos, 2018, Section 12.7, Paragraph 3).

Applying this recommendation to our data, we found that all estimated AICc values were in favor of the ARIMA models.

Software

All analyses were conducted in RStudio 1.3.959 (RStudio Team, 2020) based on R version 4.0.1 (R Core Team, 2020). For tweet collecting, we used the package rtweet 0.7.0 (Kearney, 2019), for text mining quanteda 2.0.1 (Benoit et al., 2018), for topic modeling BTM 0.3.1 (Wijffels, 2020), and for time series analysis the packages forecast 8.12 (Hyndman & Khandakar, 2008) and changepoint 2.2.2 (Killick & Eckley, 2014). The complete analysis code can be found in PsychArchives-ESM 2.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. http://doi.org/10.21105/joss.00774

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. http://www.jmlr.org/papers/v3/blei03a

Franses, P. H. (1994). Fitting a Gompertz curve. *Journal of the Operational Research Society, 45*(1), 109–113. https://doi.org/10.1057/jors.1994.11

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice: OTexts*. https://otexts.com/fpp2/

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software, 26*(3), 1–22. https://dx.doi.org/10.18637/jss.v027.i03

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in Psychology, 6*, 727. https://doi.org/10.3389/fpsyg.2015.00727

Kearney, M. W. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software, 4*(42), 1829, https://doi.org/10.21105/joss.01829

Killick, R., & Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software, 58*(3), 1–19. https://dx.doi.org/10.18637/jss.v058.i03

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri H. & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 conference on empirical methods in natural language processing* (p. 262–272). Edinburgh, Scotland, UK: Association for Computational Linguistics.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [Computer software].

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

RStudio Team. (2020). *RStudio: Integrated development for R*. Boston, MA: RStudio, Inc. [Computer software]

Tuleya L. G. (Hrsg.). (2007). *Thesaurus of psychological index terms* (11th ed.). Washington, DC: American Psychological Association.

Wijffels, J. (2020). *BTM: Biterm Topic Models for Short Text*. R package version 0.3.1. https://CRAN.R-project.org/package=BTM

Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013, May). A biterm topic model for short texts. In D. Schwalbe, V. A. Fernandes Almeida, H. Glaser, R. A. Baeza-Yates & S. B. Moon (Eds.), *Proceedings of the 22nd international conference on World Wide Web* (1445–1456). New York, NY: ACM. https://doi.org/10.1145/2488388.2488514

R Session Info

```
> sessionInfo()
R version 4.0.1 (2020-06-06)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 17763)

Matrix products: default

locale:
[1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252    LC_MONETARY=German_Germany.1252
[4] LC_NUMERIC=C                    LC_TIME=German_Germany.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] forecast_8.12      ggraph_2.0.3      ggplot2_3.3.1     textplot_0.1.2    BTM_0.3.1         udpipe_0.8.3
 [7] data.table_1.12.8 rtweet_0.7.0      changepoint_2.2.2 zoo_1.8-8         quanteda_2.0.1

loaded via a namespace (and not attached):
 [1] ggrepel_0.8.2      Rcpp_1.0.4.6       lubridate_1.7.8    lattice_0.20-41    tidyr_1.1.0        digest_0.6.25
 [7] packrat_0.5.0      lmtest_0.9-37      ggforce_0.3.1      R6_2.4.1           httr_1.4.1         pillar_1.4.4
[13] rlang_0.4.6        curl_4.3           rstudioapi_0.11    TTR_0.23-6         fracdiff_1.5-1     Matrix_1.2-18
[19] igraph_1.2.5       polyclip_1.10-0    munsell_0.5.0      compiler_4.0.1     pkgconfig_2.0.3    urca_1.3-0
[25] nnet_7.3-14        tidyselect_1.1.0   tibble_3.0.1       gridExtra_2.3      quadprog_1.5-8     graphlayouts_0.7.0
[31] viridisLite_0.3.0  crayon_1.3.4       dplyr_1.0.0        withr_2.2.0        MASS_7.3-51.6      grid_4.0.1
[37] nlme_3.1-148       jsonlite_1.6.1     gtable_0.3.0       lifecycle_0.2.0    magrittr_1.5       scales_1.1.1
[43] RcppParallel_5.0.1 quantmod_0.4.17    stringi_1.4.6      farver_2.0.3       viridis_0.5.1      fs_1.4.1
[49] tseries_0.10-47    timeDate_3043.102  xts_0.12-0         ellipsis_0.3.1     stopwords_2.0      generics_0.0.2
[55] vctrs_0.3.0        fastmatch_1.1-0    tools_4.0.1        glue_1.4.1         tweenr_1.0.1       purrr_0.3.4
[61] parallel_4.0.1     colorspace_1.4-1   tidygraph_1.2.0    usethis_1.6.1
```