Investigating the Effects of MWE Identification in Structural Topic Modelling

Dimitrios Kokkinakis
University of Gothenburg and Centre for Ageing and Health (AgeCap), Sweden
dimitrios.kokkinakis@gu.se

Ricardo Muñoz Sánchez
University of Gothenburg, Sweden
ricardo.munoz.sanchez@gu.se

Sebastianus C. J. Bruinsma
Chalmers University of Technology, Sweden
sebastianus.bruinsma@chalmers.se

Mia-Marie Hammarlin
Lund University and Birgit Rausing Centre for Medical Humanities (BRCMH), Sweden
mia-marie.hammarlin@kom.lu.se

Abstract

Multiword expressions (MWEs) are common word combinations which exhibit idiosyncrasies in various linguistic levels. For various downstream natural language processing applications and tasks, the identification and discovery of MWEs has been proven to be potentially practical and useful, but still challenging to codify. In this paper we investigate various, relevant to MWE, resources and tools for Swedish, and, within a specific application scenario, we apply structural topic modelling to investigate whether there are any interpretative advantages of identifying MWEs.

1 Introduction

Multiword expressions (MWEs) are common word combinations which exhibit idiosyncrasies on a variety of lexical, syntactic, semantic, pragmatic and/or statistical levels. In this paper we investigate the impact of multiword expression (MWE) identification as a text preprocessing step prior to the application of a structural topic modeling (STM) approach (Roberts et al., 2019). As a case scenario, for investigating the feasibility of this experimental set-up, we provide an exploratory comparison of the STM analysis on a dataset that contains Swedish social medial posts about novel vaccines (mRNA, Novavax), *with* and *without* the identification of MWEs. The aim of this work is to answer the following research question: *in a mixed-method research design can MWE identification enhance the interpretability and explainability of the generated topics and themes?*

We start by applying both available MWE lexical resources for Swedish (lexicons and processing tools) and standard extraction techniques (e.g., n-gram collocations) to preprocess the vaccine-related narratives by keeping one version of the dataset intact, i.e., without any MWEs identified. Then, we apply STM in the two versions of the same dataset, to uncover the most prevalent discussion topics. As a methodological step, we utilize an exploratory mixed quantitative-qualitative approach (Ivankova et al., 2016) to investigate, compare, discuss, and loosely evaluate these topics with respect to the specific application scenario at hand (cf. Section 2). The motivation behind the application is based on the fact that multiword expressions can improve topic coherence, which is positively correlated with human assessment and readability of topics (Aletras & Stevenson, 2013). MWEs can also be used to reduce ambiguity, for example, by recognizing multiword terms/names as opposed to single word tokens could prevent an incorrect interpretation in many domains; for instance, *autoimmun reaction* ‘autoimmune reaction’ (instead of *autoimmun* and *reaction* separately) or *vitamin D-brist* ‘vitamin D deficiency’ (instead of *vitamin* and *D-brist* separately) (cf. Spasic & Button, 2020; Kochmar et al., 2020).

2 Application scenario: Swedish social media data about novel vaccines

We use vaccine skepticism as the application scenario for our case study. Vaccine skepticism can be triggered by anxiety about possible side effects and concerns related to novel vaccine technologies, such as the messenger RNA (mRNA) which can be used as a reason for not

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 36–44

receiving (the COVID-19) vaccine (Leong et al., 2022). For instance, the University of Lund study: “Intracellular Reverse Transcription of Pfizer BioNTech COVID-19 mRNA Vaccine BNT162b2 In Vitro in Human Liver Cell Line” (Aldén et al., 2022), published on February 2022, has been frequently cited since its release, as a confirmation for vaccine skepticism and hesitancy, highlighting a potential misconception that the mRNA vaccine alters the human DNA.

2.1 Dataset: Swedish social media data on novel vaccines and vaccination

Having the aforementioned study as one of our starting points, we extracted Swedish tweets downloaded from February 10, 2022 (two weeks before the Lund study was published) to November 10, 2022 (nine months in total). The tweets were collected using the rtweet package (v1.1.0), which provides access to the Twitter API from R[1]. The final tweet data set consisted of 1,870 unique tweets from 858 different users (26,000 tokens without punctuation and with stop words removed). Furthermore, we extracted 8,900 unique social media posts (80,000 tokens; again, without punctuation and with stop words removed), from the popular Swedish forum Flashback[2]. These posts originate from Flashback vaccine-related threads published around the same period as above.

Two versions of the whole dataset were produced, the first version we name *sv-socialMedia-original* (without any labelled MWEs) and the second *sv-socialMedia-mwe* (the version with labelled MWEs; *cf.* 2.2).

2.2 Swedish MWE resources

For the MWE exploration in STM we decided to use as a preprocessing step any, as far as possible, available, Swedish resources for MWE annotation. We applied the following resources:

i. available multiword lists and lexicons with multiword entries, such as Swedish lexicalized idioms[3] (e.g., *vind i seglen* ‘wind in your sails’ [i.e., to have success] and *käppar i hjulet* ‘stand in the way’) and phrasal verbs, including inflected forms[4] (e.g., *komma ihåg* ‘to remember’ and *bryta ner* ‘break down’);
ii. named entity recognition for Swedish (Kokkinakis et al., 2014). Named entities that are composed of more than one token were kept and marked as a MWE; e.g., *Robert Malone; Falun Gong* and *Bill Gates foundation*);
iii. function words, mainly adverbs and prepositions[5] (e.g., *till följd av* ‘as a result of’ and *på grund av* ‘because of’);
iv. medical terminology, particularly names of symptoms[6] from ICD-10, and disease names. Technical and medical terms usually form non-compositional compound terms because of the need for specificity. Combining such terms into single token compounds may result in improved specificity / comprehensibility in the topics (cf. Boyd-Graber et al., 2017). E.g., *försämrat immunförsvar* ‘impaired immune system’;
v. statistically significant n-gram collocations[7] (basically bigrams and some trigrams) after manual selection of the top-400 strongest collocations. For instance, some of the highly ranked n-grams, acquired from the dataset, include: *kognitiv förmåga* ‘cognitive ability’; *fertil ålder* ‘fertile age’; *plötsligt hjärtstopp* ‘sudden cardiac arrest’; *naturlig*

[1] The keywords that were used included the pattern: (m-?[Rr][Nn][Aa]|[Nn]ovavax).* (‘?’ the preceding character is optional; ‘|’ disjunction and ‘.*’ ≥ 0 characters following) or the hashtags #mRNA / #novavax and lang:sv (Swedish).

[2] `https://www.flashback.org/`.

[3] The lexicalized idioms (ca 4,000) originate from the NEO lexicon: `https://spraakbanken.gu.se/en/resources/neo-idiom`.

[4] Collection from various Internet sources e.g., the Swedish Wiktionary `https://sv.wiktionary.org/wiki/Kategori:Svenska/Partikelverb`; and the manually annotated Swedish verbal MWEs in the 1.2 edition of the PARSEME Shared Task, particularly the categories, Verb-particle constructions (VPC.full) and the inherently reflexive verbs (IRV) `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3367`.

[5] `https://sv.wiktionary.org/wiki/Appendix:Svenska_flerordiga_prepositioner`.

[6] ICD-10-SE: International Statistical Classification of Diseases and Related Health Problems 10th Revision. R00-R99: Symptoms, signs and abnormal clinical and laboratory findings: `https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/icd-10/`.

[7] Using the R package `quanteda` (v. 0.9.9-65).

immunitet 'natural immunity' and *villkorat godkännande* 'conditional authorization'.

2.3 Data Preprocessing

Data preprocessing is a critical step in the raw text analysis process. Maier et al. (2018) emphasize the fact that appropriate preprocessing of the text collection is one of the major challenges researchers need to tackle for topic modeling application to textual data. This procedure involves a series of actions to clean and normalize text with the goals of removing potential noise and consequently obtain a better quality of the data and the topics for the dataset (cf. Section 2.1). The *sv-socialMedia-mwe* dataset is preprocessed with the annotation of the resources described in Section 2.2, in which multiword tokens are concatenated by an underscore character to a single token for uniformity (e.g., *Robert_Malone*). However, before annotation, both versions of the social media textual content were normalized. Basically, letters were converted from uppercase to lowercase; punctuations were stripped off and stop-words were removed (for the *sv-socialMedia-mwe* stop-words were removed after the MWE recognition).

In addition to the standard stop-words, we stripped off the top-10 most frequent and corpus-specific words, e.g., *vaccin* 'vaccine'; *vaccinera* 'to vaccinate'; *biverkning* 'side effect' and *sjukdom* 'disease'. These are repeated words that negatively affect the quality of topic models, leaving less representational power for the remaining text and, consequently, most likely yield less coherent topics (Almgerbi et al., 2021). As in other studies (cf. Duraivel & Lavanya, 2021) we neither stemmed nor lemmatized the dataset since stemming can alter the context of some of the words important for model building and interpretability. Moreover, we didn't lemmatize since we wanted to also investigate whether inflection could bring some valuable interpretative information. We are aware that this is a threshold hard to meet since lemmatization can also dilute useful information regarding a single concept into several inflected forms. To avoid missing out important information, we kept the corpus unstemmed and unlemmatized for the analysis. The identified MWEs during the preprocessing phase are replaced with single tokens before running the SMT, which does not induce any additional complexity to the models used in SMT. The distribution of the MWE types (relative frequency) is shown in Figure 1; moreover, in absolute values there were 3,926 bigrams; 628 trigrams and 26 tetragrams in the data.

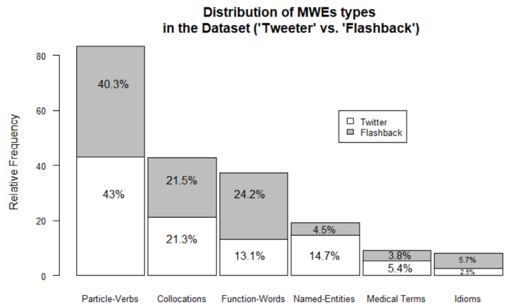


Figure 1: Distribution of the relative frequency of the MWE types in the dataset.

3 Related work

Latent Dirichlet Allocation (LDA) (Blei et al., 2002), is a popular topic modeling method that uses the statistical analysis of textual data to identify themes or topics that occur in a document collection. Although topic modelling can identify the topics contained in text, the original bag of words approach used by LDA models ignores the order of the words which limits the deeper understanding of the content. Therefore, during the last years researchers try to enhance the various flavors of topic models with the addition of n-gram features to improve the results and reduce the complexity of the models. Particularly, phrase-based topic modeling has shown significant improvement, especially on short text data (Kherwa & Bansal, 2020; Nokel & Loukachevitch, 2016); and several studies acknowledge the fact LDA results can be improved when MWE expressions are included during processing (Wang et al., 2016; Guarino & Santoro 2018; Cheevaprawatdomrong et al., 2022).

4 Structural Topic Modeling

Structural topic model (STM) has emerged as an extension to LDA allowing the integration of covariates into the prior distributions for document-topic compositions and topic-word proportions. Thereby, STM can be used to model how the content of a collection of documents changes as a function of document-level covariates such as day and time, and gain insights and understanding on how topics evolve (Lebryk 2021).

We apply STM to automatically detect latent topics in the dataset which can be used to investigate the nature of these topics reflected in the novel vaccine discussions (Scannell et al., 2021). We use a sequential explanatory mixed methods approach which consists of a quantitative phase (collection, cleaning, and natural language processing of the data), followed by a qualitative phase (in-depth analysis of the results from the quantitative phase). This type of design provides greater analysis depth than either singular analysis would (Fetters et al., 2013).

4.1 STM parameters

Since there is no "correct" solution for determining the optimal number of topics *k* that should be generated during the model selection process, several diagnostic aspects of the topic modeling were evaluated to decide the number of topics, *k*, to use. The *stm* package implements several evaluation metrics, such as the spread of *semantic coherence* (Mimno et al., 2011) and *exclusivity*, which both capture what humans qualitatively perceive as good topics (Roberts, et al., 2019). After preprocessing of the data, a document-term matrix was created and used for modeling, while the best model yielded 6 topics. This number was chosen after running multiple STM models, ranging from 2 to 40 topics (Roberts et al., 2019). We then used a combination of quantitative (exclusivity and semantic coherence) and qualitative methods to decide on the final numbers of topics (Appendix A), in order to evaluate the performance of structural topic modelling algorithm. The semantic coherence score measures the degree of semantic similarity between high-scoring words in the topic and ranges from $-\infty$ to 0. High semantic coherence measurements help distinguish between topics that are semantically interpretable while low scored topics are usually artifacts of statistical inference. Exclusivity measures the extent to which the top words for each topic do not appear as top words in other topics. Exclusivity ranges from 0 to $+\infty$. These quantitative metrics measure to what degree topics contain many overlapping words, and to what degree words that occur in the same topic also occur in the same context. For simplification reasons during the comparison between the two dataset versions, we set the number of topics to be the same. Figure 2 shows the semantic coherence vs the exclusivity of the models (40 topics), while Appendix B shows the temporal evolution of the identified topics.

5 Results and discussion

We have hypothesized that the identification of multiword expressions can provide us with better and more targeted insights and enhance the interpretability and explainability of the generated topics and themes. The characterization of the multiword expression types which are recognized and applied in this experimental setup follows the order given in Section 2.2.

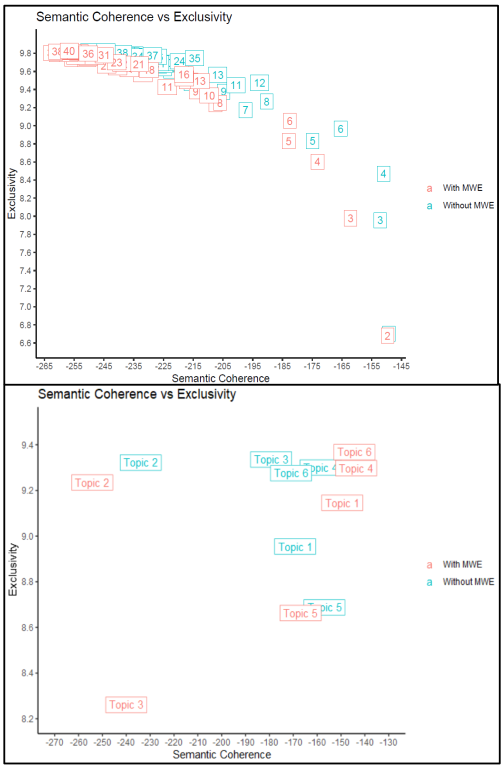


Figure 2. Comparison of the number of topics. Semantic coherence vs exclusivity of the generated models in the 2 versions of the dataset (top fourty topics and bottom six).

We extracted a list of keywords for each topic that have the highest association with that certain topic. For this association, we base ourselves on the FREX (frequency and exclusive) value of each word. This value combines the exclusivity of each word (meaning that a word occurs more often in that topic than in others) while also correcting for its overall frequency (Airoldi and Bischof 2016). To better understand and label the topics and to label them, we also extracted the top

30 most representative posts for each topic.

This qualitative evaluation shows that the proposed method provides only *slightly* better performance of SMT on the *sv-socialMedia-mwe* dataset. Moreover, the FREX metric helps us to assign intuitive labels to the subject matter. FREX is defined as the ratio of word frequency, and subject to word-topic exclusivity. Balancing these two measurements is important as frequent words can often be uninformative, while completely exclusive words can be very rare and not informative. The FREX example below, for instance, taken from the *sv-socialMedia-mwe* version (of topic 2 'women's health issues'), illustrates that three of the tokens, among the top 12, are multiwords: *mensrubbning* 'menstrual disorder'; *polio* 'polio'; *missfall* 'miscarriage'; *fertilitet* 'fertility'; *graviditet* 'pregnancy'; *stelkramp*, 'tetanus'; *klimakteriet* 'menopause'; *rubbning* 'disturbance'; *[gravid_kvinna]* 'pregnant woman'; *menstruation* 'menstruation'; *[röda_hund]* 'rubella'; and *[fertil_ålder]* 'fertile age'.

6 Conclusions and Future Work

The results suggest that the STM models perform only s*lightly* better without MWE than with. One reason could be that the vocabulary is larger without MWEs included, so the number of topics capture more words. Still, the differences are quite small, and the comparisons of the models based on the two datasets are not 100% fair, since the MWE accounts for about 600 "terms" more and is therefore kind-of another dataset altogether as far as the STM is concerned.

Still, there is some indication that MWE identification leads to better interpretability of the STM, as calculated by the semantic coherence, which was higher for the dataset with MWEs. Therefore, we could conclude that the identification of MWEs can slightly enhance the explainability of the generated topics and themes, which could lead to a more appropriate labeling of the topic itself during the qualitative interpretation of the generated topics, i.e., incorporating multiword expressions into the models, creates slightly more informative resulting topics.

In general, the major differences between the two versions of the topics are also shown in the graphs of Appendix B which show the prevalence over time for the topics of the two dataset versions – without stating anything about their quality. Major differences for some of the generated topics, that needs further investigation.

As a future work it would be also interesting to verify the efficacy of our resources and our method on different domains and types of datasets and explore more resources for multiword recognition for Swedish. Identification of non-contiguous multiword expressions is another area we need to explore (Barreiro & Batista, 2016). There are several opportunities for future research to extend our assessment of the performance and evaluation. As previous research has pointed out, topic modeling algorithms are sensitive to several characteristics such as text length and the text preprocessing applied, for instance no stemming or lemmatization was applied that could have impact on the results (Stoy, 2021; Rüdiger et al., 2022). Hence, further investigations including parameters and characteristics are necessary.

Limitations

Since the extraction of the dataset content in this paper is based on a rather polarized dataset and from only two sources, future analyses will focus on testing the reliability of this research on other (larger) text collections. Furthermore, the dataset size is rather small; a limitation of the presented work is the search itself which only used a non-exhaustive list of keywords, basically on novel vaccines and vaccination. Moreover, we have no clue on how diverse the socio-demographic backgrounds of the users are, and therefore how commenting could be related to different sociodemographic characteristics of the sample. Finally, the multiword expressions we explored were all *contiguous*, results would probably be more comprehensive and profound if non-contiguous multiword expressions could also be identified and modelled. Similarly, lemmatization could be an important step to further explore since many of the generated clusters included inflected variants of the same word.

Acknowledgments

This work is part of the project Rumour Mining (MXM19-1161:1) financed by the Bank of Sweden Tercentenary Foundation (Riksbankens Jubileumsfond).

References

Edoardo Airoldi and Jonathan M. Bischof. 2016. Improving and Evaluating Topic Models and Other Models of Text. J of the American Statistical Association 111 (516): 1381-1403. DOI: 10.1080/01621459.2015.1051182.

Markus Aldén, Francisko Olofsson Falla, Daowei Yang, Mohammad Barghouth, Cheng Luan, Magnus Rasmussen and Yang De Marinis. 2022. Intracellular Reverse Transcription of Pfizer BioNTech COVID-19 mRNA Vaccine BNT162b2 In Vitro in Human Liver Cell Line. *Curr. Issues Mol. Biol.* 44(4), 1661-1663.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*. Pp. 13-22. Potsdam, Germany. Association for Computational Linguistics.

Mohamad Almgerbi, Andrea De Mauro, Adham Kahlawi and Valentina Poggioni. 2021. Improving Topic Modeling Performance through N-gram Removal. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 162-169).

Anabela Barreiro and Fernando Batista. 2016. Machine Translation of Non-Contiguous Multiword Units. *Proceedings of the Workshop on Discontinuous Structures in Natural Language Processing.* Pp 22-30. 10.18653/v1/W16-0903.

Jonathan M. Bischof and Edoardo M. Airoldi. 2012. Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning. ICML*. Pages 9–16

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3. 993-1022.

Jordan Boyd-Graber, Yuening Hu and David Mimno. 2017. *Applications of Topic Models*. Foundations and Trends® in Information Retrieval. Vol. 11:2-3, pp 143-296. 10.1561/1500000030.

Jin Cheevaprawatdomrong, Alexandra Schofield and Attapol Rutherford. 2022. More Than Words: Collocation Retokenization for Latent Dirichlet Allocation Models. *Proceedings of the ACL Workshop: Findings of the Association for Computational Linguistics*. Pp 2696-2704. Dublin, Ireland 10.18653/v1/2022.findings-acl.212.

Leong Ching, Jin Lawrence, Kim Dayoung, Kim Jeongbin, Teo Yik Ying and Ho Teck-Hua . 2022. Assessing the impact of novelty and conformity on hesitancy towards COVID-19 vaccines using mRNA technology. *Commun Med.* 2:61.

Samuel Duraivel and Aby Augustine Lavanya. 2021. Understanding vaccine hesitancy with application of latent dirichlet allocation to reddit corpora. *Ind J of Science and Technology*. Vol 15:37, pp. 1868-1875. 10.21203/rs.3.rs-616664/v1.

Michael D. Fetters, Leslie A. Curry and John W. Creswell. 2013. Achieving Integration in Mixed Methods Designs. Health Serv. Res. 48, 2134-2156. 10.1111/1475-6773.12117.

Stefano Guarino and Mario Santoro. 2018. Multi-word Structural Topic Modelling of ToR Drug Marketplaces. *IEEE 12th International Conference on Semantic Computing (ICSC)*. CA, USA, pp. 269-273. 10.1109/ICSC.2018.00048.

Nataliya V. Ivankova, John W. Creswell and Sheldon L. Stick. 2016. Using mixed-methods sequential explanatory design: From theory to practice. *Field methods.* 18(1):3-20. 10.1177/1525822X05282260.

Pooja Kherwa and Poonam Bansal. 2020. Semantic N-gram topic modeling. *EAI Endorsed Transactions on Scalable Information Systems*, 7 (26). 10.4108/eai.13-7-2018.163131.

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. Detecting Multiword Expression Type Helps Lexical Complexity Assessment. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (LREC). Pp 4426–4435, Marseille, France. European Language Resources Association.

Dimitrios Kokkinakis, Jyrki Niemi, Sam Hardwick, Krister Lindén, and Lars Borin. 2014. HFST-SweNER — A New NER Resource for Swedish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2537–2543, Reykjavik, Iceland. European Language Resources Association (ELRA).

Theo Lebryk. 2021. Introduction to The Structural Topic Model (STM). A unique way to use topic modelling for social science research. Towards Data Science: https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383

Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri and S. Adam. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Com Methods & Measures*. 12:2-3, 93-118, 10.1080/19312458.2018.1430754.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum 2011. Optimizing Semantic Coherence in Topic Models.

Proceedings of the Empirical Methods in NLP. Edinburgh, UK, c Association for Computational Linguistics, p. 262–272.

Michael Nokel and Natalia Loukachevitch. 2016. Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 44–49, Berlin, Germany, August 7-12, 2016. Association for Computational Linguistics.

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley. 2019. Stm: An R package for structural topic models. *Journal of Statistical Software.* 91(2), 1–40. 10.18637/jss.v091.i02.

Matthias Rüdiger, David Antons, Amol M. Joshi and Torsten-Oliver Salge. 2022. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLoS One*. 17(4):e0266325. 10.1371/journal.pone.0266325.

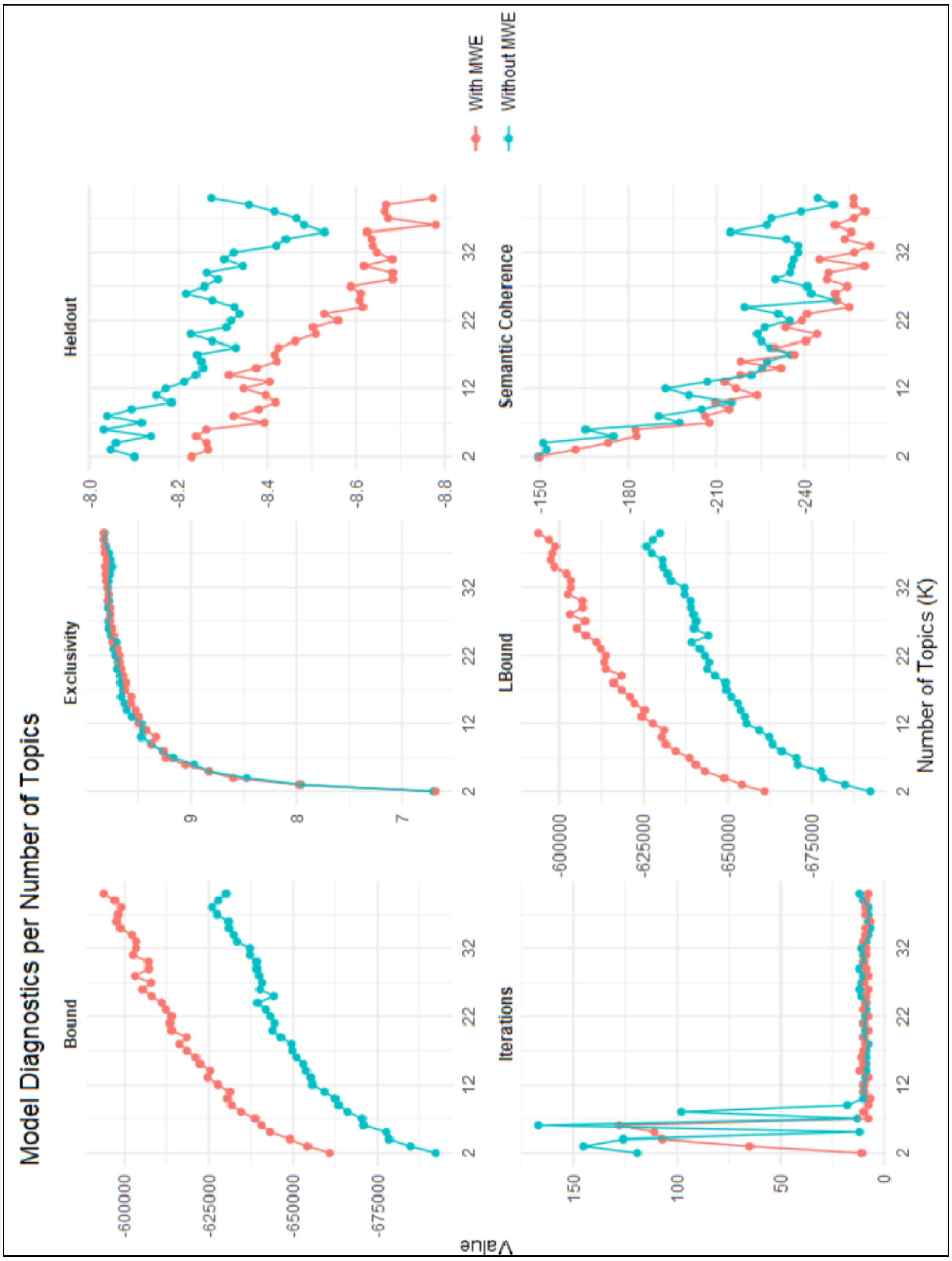
Denise Scannell, Linda Desens, Marie Guadagno, Yolande Tra, Emily Acker, Kate Sheridan, Margo Rosner, Jennifer Mathieu and Mike Fulk. 2021. COVID-19 Vaccine Discourse on Twitter: A Content Analysis of Persuasion Techniques, Sentiment and Mis/Disinformation. *J Health Com* 3;26(7). 10.1080/10810730.2021.1955050.

Irena Spasic and Kate Button. 2020. Patient Triage by Topic Modeling of Referral Letters: Feasibility Study. *JMIR Med Inform.* 6;8(11):e21252. 10.2196/21252.

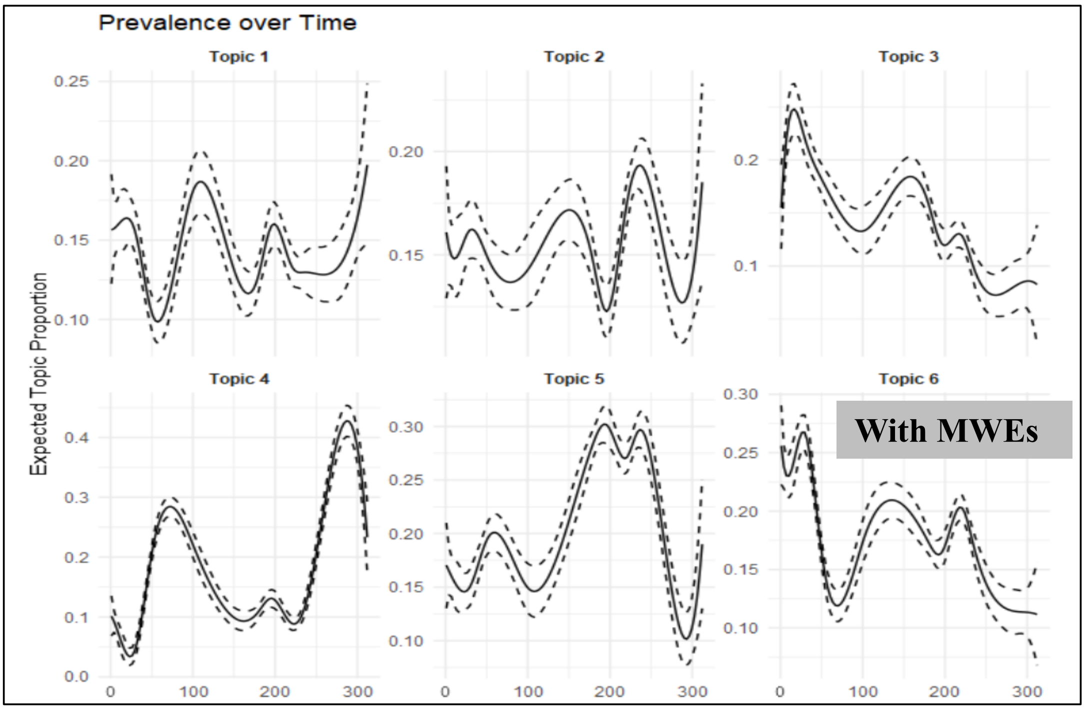
Lazarina Stoy. 2021. *8. Limitations of Topic Modelling Algorithms on Short Text*. https://lazarinastoy.com/topic-modelling-limitations-short-text/ (visited 2023-02-11).

Minmei Wang, Bo Zhao, and Yihua Huang. 2016. Ptr: phrase-based topical ranking for automatic keyphrase extraction in scientific publications. *International Conference on Neural Information Processing.* Pp 120–128. Springer. 10.1007/978-3-319-46681-1_15.

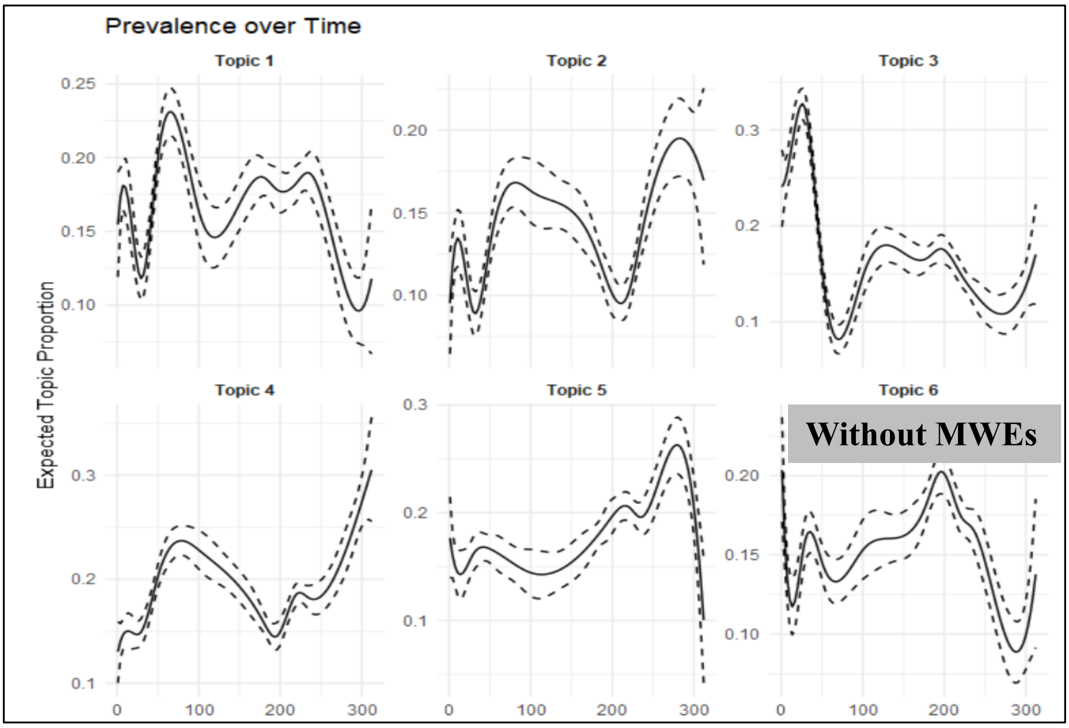
Appendix A: Model Diagnostics.



Appendix B: Topic prevalence over time.



Timeline: February-November 2022



Timeline: February-November 2022