Understanding Narratives from Demographic Survey Data: a Comparative Study with Multiple Neural Topic Models

Xiao Xu
NIDI-KNAW
University of Groningen
xu@nidi.nl

Gert Stulp
University of Groningen
g.stulp@rug.nl

Antal van den Bosch
Utrecht University
a.p.j.vandenbosch@uu.nl

Anne Gauthier
NIDI-KNAW
gauthier@rug.nl

Abstract

Fertility intentions as verbalized in surveys are a poor predictor of actual fertility outcomes, the number of children people have. This can partly be explained by the uncertainty people have in their intentions. Such uncertainties are hard to capture through traditional survey questions, although open-ended questions can be used to get insight into people's subjective narratives of the future that determine their intentions. Analyzing such answers to open-ended questions can be done through Natural Language Processing techniques. Traditional topic models (e.g., LSA and LDA), however, often fail to do since they rely on co-occurrences, which are often rare in short survey responses. The aim of this study was to apply and evaluate topic models on demographic survey data. In this study, we applied neural topic models (e.g. BERTopic, CombinedTM) based on language models to responses from Dutch women on their fertility plans, and compared the topics and their coherence scores from each model to expert judgments. Our results show that neural models produce topics more in line with human interpretation compared to LDA. However, the coherence score could only partly reflect on this, depending on the method and corpus used for calculation. This research is important because, first, it helps us develop more informed strategies on model selection and evaluation for topic modeling on survey data; and second, it shows that the field of demography has much to gain from adopting novel NLP methods.

1 Introduction

Demographers are interested in the number of children people have or will have, also referred to as fertility. In trying to understand future fertility, researchers have studied fertility intentions, i.e. plans to have children in the future. The usefulness of measurements of fertility intentions are often debated among demographers due to the gap between intentions and fertility outcomes (Brinton et al., 2018; Trinitapoli and Yeatman, 2018) and a large portion of respondents being uncertain about their intentions (Bhrolcháin and Beaujouan, 2019). It is proposed that this is because fertility intentions are contextual and largely depend on subjective narratives (Vignoli et al., 2020). Therefore, understanding these narratives might be the key for advancing theories on the fertility decision-making process.

Open-ended questions (OEQs) help researchers obtain "top-of-the-head" answers from respondents, and they have been employed in previous qualitative demographic studies (e.g., interviews with a small sample of respondents, sometime deliberately non-representative of the whole population) (Schatz and Williams, 2012; Staveteig et al., 2017). However, to expand the analysis to a larger and generalizable sample of the population, an automatic process of extracting and quantifying themes from responses is needed as an initial exploratory data analysis. This objective can be met with topic modeling methods.

Latent Dirichlet Allocation (LDA; Blei et al. (2003)) is one of the most popular topic modeling algorithms, which is based on co-occurrence of words. LDA's performance on short texts, such as online survey responses, may be compromised due to the small number of co-occurrences. To overcome this problem, many topic models that support incorporating prior language knowledge (e.g. word embeddings or language models) have been developed, such as *Sparse Contextual Hidden and Observed Language AutoencodeR* (SCHOLAR; (Card et al., 2017)). This model uses variational autoencoder (VAE) to incorporate word2vec (Mikolov et al., 2013) embeddings. A different example is Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. (2018)), one of the most prominent language models, that has also been incorporated in topic modeling tasks, e.g. in Combined Topic Model (CombinedTM; Bianchi

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 33 - 38

et al. (2020)) and BERTopic (Grootendorst, 2022).

In this paper, our first aim is to compare the performance of multiple topic modelling algorithms on responses to open questions. Such an analysis of survey responses is rare, and it is an open question how well these topic models do on short texts from relatively few respondents (400), a scale larger than usual qualitative studies. To achieve this, we implemented and evaluated four models (LDA, SCHOLAR, CombinedTM and BERTopic), trained on the fertility response dataset to provide unsupervised topics. We then compare metrics of quality across these diverse methods through comparable implementations. Building on the comparative study of Baumer et al. (2017), we further compare metrics and their difference to human annotations respectively.

Our study contributes to the literature by: 1) evaluating the performance of multiple topic models incorporating prior knowledge; 2) examining the correlation between metrics and human judgments; 3) modeling topics on Dutch texts of online survey responses and 4) bringing novel text analysis methods to the field of demography.

2 Data

The data used in this study is collected through the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata at Tilburg University, The Netherlands. The study is based on the second wave of survey *Social networks and fertility (in Dutch: Sociale relaties en kinderkeuzes onderzoek)* within the LISS panel in 2021, which was first fielded in 2018. The module's objective was to investigate fertility intentions and attitudes in relation to people's personal networks. For this round of our survey, 596 female participants were invited, and 464 women between the ages of 21 and 44 completed the questionnaire. The survey was conducted in Dutch. The open-ended questions (henceforth: OEQ) regarding fertility intentions are presented to respondents that are not currently pregnant (N=433). After removing 6 answers that were without information (e.g. "niets") or not in Dutch, there were in total 427 responses available.

The OEQ was placed directly after a standard closed questions on fertility intention ("Do you intend to have a/another child during the next three years?") from the Generations & Gender Surveys (GGS) (Gauthier et al., 2018). Respondents were presented with a text box, where they can input text answers. Two versions of the OEQ were tested:

- **Original** Can you tell us more about what makes you (un)certain about whether or not to have children?
- **Adaptive reminder** You answered the previous question "Do you plan to have a child in the next three years?" with [*[1]]. Can you tell us more about what makes you (un)certain about whether or not to have children?

The answers contain 32 words on average. Since answers to these two questions were similar on a suite of textual characteristics (e.g., sentence length, number of nouns), we did not differentiate between answers to these two questions in the subsequent analyses.

3 Evaluation

Each model was evaluated on three different metrics: topic coherence, topic diversity, and comparison to human-assigned labels.

Topic coherence measures how close the top n words (typically, $n = 10$) from a topic are to each other: if the words always co-occur in documents, they are considered "close" and the topic is considered *coherent*. It is calculated through non-negative point-wise mutual information (NPMI, Newman et al. (2010)), where w denotes a word:

$$\text{NPMI}(w_i) = \sum_{j}^{n-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_j,w_j)}$$

The calculation of topic coherence requires an external test corpus for calculating how frequently words in the topic occur together in real language usage Blair et al. (2020). The external corpus was crawled from the Viva forum, a Dutch online discussion board mainly aimed at women. We prepared two corpora for cross validation: first the "Child wish" corpus, which contains 436 threads and replies including the keyword "child wish[2]"; second the "Pregnancy[3]" corpus, containing 5507 comments under the "Pregnancy" board. Coherence scores were calculated on both corpora.

Topic diversity measures how different the top ten words from all topics are; i.e. if topics share

[1] definitely yes / probably yes / unsure / probably not / definitely not/ don't know

[2] In Dutch: *kinderwens*

[3] In Dutch: *Zwanger*

the same words. It is calculated through Inversed Rank-Biased Overlap (ρ; Webber et al. (2010)), where the top ten words are compared. The score ranges from 0 to 1, representing topics that contain exact same words to totally different topics.

The four authors of this paper intensively read on recent fertility trends and events in the Netherlands. The first and fourth authors read through the dataset to develop qualitative insights and proposed themes of discussions from respondents; as we coded the responses iteratively, although exact labels were not yet given to each response, six themes were summarized. Then, all authors interpreted lists of top words for each topic generated by models, and compared results with different number of topics K to develop the ideal K. Eventually, we have together determined the optimal number of topics to be $K = 9$ and established a verbal theme for each topic.

We ran grid search on each model for hyper-parameter tuning under the same number of topics $K = 9$; this did not apply to BERTopic as it used HDBSCAN algorithm and did not require a pre-defined K. We use COW word embedding (Tulkens et al., 2016) for SCHOLAR model, and we use RobBERT (Delobelle et al., 2020), a state-of-the-art Dutch BERT model for CombinedTM and BERTopic.

4 Results

In this section, we first present the themes from qualitative insights. These are then compared to results from the four topic models. Since human produced themes may correspond to more than one topic (Baumer et al., 2017), we calculate how many themes were accounted for and present them together with other metrics. The themes and corresponding topics are presented in Table 1. Each relevant topic has a human-assigned label, describing its perceived content; if there are multiple topics relevant to one theme, they are separated by the & symbol.

4.1 Qualitative insights

These insights were summarized by the authors of this paper through rounds of reading and discussion. Here, people talked about the issues and conditions about what made them feel uncertain about having kids. Age and family size were the most prominent themes, while various other personal circumstances and societal issues were also mentioned.

4.1.1 Age

Age is one of the most mentioned themes in the answers; in fact, some respondent only left one word "age" in their answer. Other more elaborated responses can be divided into two groups: "too young" (e.g. "I'm only 23 years old and still studying") and "too old" (e.g. "I'm already 43 and I do not have wish for child").

4.1.2 Number of kids

Many respondents who already have kids and are satisfied with their current family size. For example, "My family is complete, and we are satisfied with 2 kids".

4.1.3 Lifestyle

This theme concerns those who have other plans or want to do a lot of things before having children, e.g. studying, traveling, finding a part-time job. It sometimes co-occurs with young age. An example is, "I would really like to have children, but at the moment I am still at an age where I also want to have time with my boyfriend to make beautiful trips and have time for the two of us".

4.1.4 Pre-conditions

Having children may require a lot of pre-conditions and this is used as justification for not wanting to have (more) children, especially among younger respondents. Conditions that were mentioned including having a stable partner, a stable job, a property, or finishing studies. This is a typical response from a student: "I will finish my studies this year, after that I first want to be able to work full-time for a number of years in order to possibly also buy a house".

4.1.5 Health issues

In our study, the theme of health issues refers to cases where the respondent wanted to or had been trying to have kids, but failed to or refrain from getting pregnant due to infertility or other medical conditions. For example, one mentioned that "the risk of complications with myself is quite high. In addition, I take medicines that are not possible in combination with a pregnancy".

4.1.6 Dissatisfaction

There is a small set of responses that, instead of personal circumstances, talked about broader dissatisfaction with the world or society. Issues raised include environmental concerns ("The world is not a nice place now. Climate changes are becoming

Theme	LDA	SCHOLAR	CombinedTM	BERTopic
1. Age	**Yes**	No	**Yes** (too young & too old)	**Yes** (too old)
2. Number of kids	No	No	**Yes** (family complete)	**Yes** (family complete)
3. Lifestyle	No	**Yes** (sacrifice to make)	**Yes** (early stage of life)	**Yes** (freedom)
4. Pre-conditions	**Yes** (partner)	**Yes** (housing & relation)	**Yes** (studies & jobs)	**Yes** (partner & jobs)
5. Health issues	No	**Yes** (health & medicine)	**Yes** (infertile & illness)	**Yes** (postnatal)
6. Dissatisfaction	No	**Yes** (climate change)	**Yes** (general)	**Yes** (economy)
Themes covered	2	4	**6**	**6**
Topic coherence (internal)	0.055	**0.464**	0.110	0.134
Topic coherence (corpus "pregnancy")	0.134	0.050	0.096	**0.158**
Topic coherence (corpus "child wish")	0.110	0.052	0.098	**0.146**
Topic diversity	0.506	**1**	0.871	0.755

Table 1: Comparison of performance between the four topic models

more and more intense"), religion ("from a biblical perspective I think it is important that I have children and hopefully let them participate in the faith"), social media ("kids nowadays are easily influenced by social media, internet, etc."). Although the issues were different by themselves, these types of responses were unified by a general sense of dissatisfaction ("I think it is very difficult to raise children in this society and world in which we now live").

4.2 Comparing topic models

We describe and compare the performance of topics in terms of each model through two sets of criteria. First, we compare topics generated by the algorithm to human produced themes, and count the number of themes that are resonated with at least one topic; then, the above-mentioned three metrics are also calculated. All results are summarized in Table 1.

We note that the two topic models that are based on BERT (CombinedTM and BERTopic) matched all themes from qualitative insights, while LDA and SCHOLAR failed to do so. This suggests that their results are closer to human judgments. However, this is only partly reflected by metrics. The SCHOLAR model, based on autoencoder and word embeddings, scored far beyond others in internal topic coherence, while BERTopic scored better than the other models on external topic coherence. SCHOLAR topped the ranking of topic diversity, while LDA scored poorly at 0.506.

Overall, we found that neural topic models indeed brought improvements over LDA: all three other models exceeded LDA by most metrics, while BERTopic outperformed LDA by all criteria.

5 Discussion

Demographers have long been calling for empirical evidence on fertility intention uncertainty and narratives (Bhrolcháin and Beaujouan, 2019; Vignoli et al., 2020). With the results from this study, we showed that neural topic models were able to provide insights similar to human judgments, thus providing a powerful tool for future demographic studies.

Our results also demonstrated the significant improvement of performance that neural topic models brought to text analysis on short survey data. The prior knowledge of language, incorporated by language models such as BERT, enabled results of quality close to traditional qualitative analysis in social studies, while previously used models such as LDA failed to so. This may have a direct applica-

tion in processing online open surveys or interview data, and enabling qualitative analysis on a larger scale.

The contrast between the ranking of scores in internal and external coherence revealed that the evaluation strategy on topic models may need to be reconsidered, especially in social science studies. Although topics generated by SCHOLAR showed an extremely high internal coherence, a closer look showed that it is mostly due to some topics consisting of words that were almost exclusively from one document (response), dragging coherence of that topic up to almost 1 (i.e. words would always co-occur). This also explained its unusually high topic diversity (at 1, which entails no repeated words at all across topics), as the topics consist of only unique words from the one document.

Our results remind us that some metrics for topic models may be misleading on a smaller, shorter dataset, and choosing the right, field-relevant corpus is a key step in correctly evaluating topic coherence. Moreover, using multiple criteria help us to avoid pitfalls and making more informed choices in selecting topic models for survey data.

Limitations

Due to limitations in time and resources, we did not conduct a thorough, full-scale grounded analysis on the corpus, as Baumer et al. (2017) did. Instead, a more lightweight approach to develop qualitative insights were chosen. Therefore, our qualitative insights and labels may still have space to improve, and the themes we proposed cannot be interpreted as a "gold standard" of model performance.

We only applied a few among many neural topic models in this study, based on easiness of implementation and availability of Dutch resources. There are several other neural topic models that are optimized for short texts, which are well summarized by Zhao et al. (2021). It would be interesting to add them for further comparisons in the future.

Ethics Statement

Ethical permission for the study was obtained from the ethical committee of sociology at the University of Groningen (ECS-201123). The dataset will be made available at `dataarchive.lissdata.nl`.

References

Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Máire Ní Bhrolcháin and Éva Beaujouan. 2019. Do people have reproductive goals? constructive preferences and the discovery of desired family size. In *Analytical family demography*, pages 27–56. Springer.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Stuart J Blair, Yaxin Bi, and Maurice D Mulvenna. 2020. Aggregated topic models for increasing social media topic coherence. *Applied Intelligence*, 50(1):138–156.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Mary C Brinton, Xiana Bueno, Livia Oláh, and Merete Hellum. 2018. Postindustrial fertility ideals, intentions, and gender inequality: A comparative qualitative analysis. *Population and Development Review*, 44(2):281–309.

Dallas Card, Chenhao Tan, and Noah A Smith. 2017. Neural models for documents with metadata. *arXiv preprint arXiv:1705.09296*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anne H Gauthier, Susana Laia Farinha Cabaço, and Tom Emery. 2018. Generations and gender survey study profile. *Longitudinal and Life course studies*, 9(4):456–465.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Enid Schatz and Jill Williams. 2012. Measuring gender and reproductive health in africa using demographic and health surveys: the need for mixed-methods research. *Culture, health & sexuality*, 14(7):811–826.

Sarah Staveteig, Richmond Aryeetey, Michael Anie-Ansah, Clement Ahiadeke, and Ladys Ortiz. 2017. Design and methodology of a mixed methods follow-up study to the 2014 ghana demographic and health survey. *Global health action*, 10(1):1274072.

Jenny Trinitapoli and Sara Yeatman. 2018. The flexibility of fertility preferences in a context of uncertainty. *Population and Development Review*, 44(1):87.

Stephan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Daniele Vignoli, Giacomo Bazzani, Raffaele Guetto, Alessandra Minello, and Elena Pirani. 2020. Uncertainty and narratives of the future: a theoretical framework for contemporary fertility. In *Analyzing contemporary fertility*, pages 25–47. Springer.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.