

# Stance-taking in topics extracted from vaccine-related tweets and discussion forum posts

Maria Skeppstedt<sup>1,2</sup>, Manfred Stede<sup>2</sup>, Andreas Kerren<sup>1</sup>

<sup>1</sup>Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden  
{maria.skeppstedt, andreas.kerren}@lnu.se

<sup>2</sup>Applied Computational Linguistics, University of Potsdam, Potsdam, Germany  
stede@uni-potsdam.de

## Abstract

The occurrence of stance-taking towards vaccination was measured in documents extracted by topic modelling from two different corpora, one discussion forum corpus and one tweet corpus. For some of the topics extracted, their most closely associated documents contained a proportion of vaccine stance-taking texts that exceeded the corpus average by a large margin. These extracted document sets would, therefore, form a useful resource in a process for computer-assisted analysis of argumentation on the subject of vaccination.

## 1 Introduction

Opinions towards vaccination that are expressed in discussion forums and in social media, as well as frequently occurring arguments given in support of these opinions, might help us to better understand reasons behind vaccine hesitancy.

There are previous studies in which such texts have been manually analysed (Grant et al., 2015; Faasse et al., 2016), as well as studies in which topic modelling has been applied for analysing texts about vaccination (Tangherlini et al., 2016; Surian et al., 2016; Skeppstedt et al., 2018).

Through topic modelling, it is possible to automatically extract topics that occur frequently in a text collection. For topic modelling to be a useful strategy for mining text collections for frequently occurring arguments, however, at least some of the topics extracted must correspond to stance positions or arguments given for these positions.

The aim of this study is to investigate if topic modelling is suitable for extracting arguments from two types of document collections that consist of laymen-produced texts about vaccination. We, therefore, measured the occurrence of stance-taking towards vaccination in the documents that were most closely associated with automatically extracted topics from two different corpora.

## 2 Background

There are previous studies that use topic modelling in computer-assisted processes to find frequently occurring arguments in a document collection (Sobhani et al., 2015; Skeppstedt et al., 2018). Documents that had been manually annotated as not containing argumentation/stance-taking were, however, removed in those two previous studies, i.e., no evaluation of the effects of including neutral documents when performing topic modelling was carried out. For most types of document collections, it is not known beforehand in which documents a stance towards the target of interest is taken or not. Therefore, the setting used here is more widely applicable, i.e., to use topic modelling on an entire text collection, without removing documents in which no stance is taken. In both of these two previous studies, the topic modelling algorithm NMF (Lee and Seung, 2001), i.e., Non-negative Matrix Factorisation, was shown appropriate for extracting arguments from short argumentative texts. We, therefore, used this algorithm in our experiments.

## 3 Method

We used topic modelling to automatically extract important topics from two different vaccination corpora, both consisting of English text that predominantly had been written by people without a medical background. We, thereafter, measured the proportion of stance-taking texts among the texts that were most related to these topics, and compared it to the proportion of stance-taking texts in the entire corpus.

### 3.1 Document collections

As a proxy for texts containing arguments, we used texts in which stance is expressed, since such texts are likely to also contain a motivation for the

position taken. The documents, from each of the two corpora, were divided into two groups based on whether they had been annotated as taking a stance towards vaccination or not, i.e., into the two groups *stance-taking* and *non-stance-taking*.

The first collection consists of posts from discussion threads on the topic of vaccination (Skeppstedt et al., 2017) that contain at least one of the following character combinations: “vacc”, “vax”, “jab”, “immunis”, and “immuniz”. Posts annotated as taking a stance *for* or *against* vaccination were combined into the group *stance-taking* texts, and posts annotated as *undecided* were assigned the category *non-stance-taking*.

The second collection consists of tweets containing the HPV vaccine-related keywords “HPV”, “human papillomavirus”, “Gardasil”, and “Cervarix” (Du et al., 2017). We combined tweets annotated according to the categories *Positive* and *Negative* to form the category *stance-taking* tweets, and tweets annotated as *Neutral* and *Unrelated* as *non-stance-taking* tweets.

Before applying topic modelling, the following were removed from the texts: standard English stop words, the terms that had been used for gathering the documents, hash tags, user names, URLs and links. Duplicated and near-duplicated documents were also removed from the collections. Documents with identical spans of texts that consisted of more than eight consecutive tokens were counted as near-duplicates. For documents consisting of ten or fewer tokens, a shorter (proportional to the length) cut-off was instead applied for classifying two documents as near-duplicates.

### 3.2 Applying topic modelling

Separate topic models were constructed for the two document collections, using the NMF class from scikit-learn (Pedregosa et al., 2011). For each topic extracted by the NMF model, the corresponding terms and documents associated with the topic are given as output, as well as their level of association with the topics.

The output of the NMF algorithm is non-deterministic, typically generating slightly different topics when run several times. Therefore, to achieve more reliable results, we followed an approach, for instance used by Baumer et al. (2017), in which the algorithm is re-run several times and only topics that occur in the output from all re-runs are retained. Before checking which topics

occurred in all re-runs, potential outliers were removed from the set of outputs from the re-runs.

We ran the algorithm 100 times with the setting to, for each re-run, return a term set consisting of the 50 terms most closely associated with each of the topics extracted by the algorithm. A topic was counted as stable when there was at least a 70% overlap between the pairs of term sets returned for a topic, for all 90 retained re-runs of the algorithm.

Potential outliers among the outputs were determined by measuring the average term overlap between the re-run outputs. That is, for each re-run, one combined set consisting of all terms associated with all topics from this re-run was constructed. Thereafter, the average overlap between this combined term set and the corresponding sets from the other re-runs was measured, i.e., the combined term sets constructed in the same fashion for each one of the other re-runs. The outputs from the 10% of the re-runs that had the lowest overlap were discarded as potential outliers, and were thus not included when calculating the stability of the extracted topics.

To avoid having to decide on a fixed number of topics in advance, which is normally required from an NMF user, we started by requesting the algorithm to extract 20 topics, and thereafter gradually decreased the number of topics requested until a maximum of 25% of the extracted topics were discarded as non-stable.

## 4 Results

After the near-duplicate filtering were 1,108 and 2,250 documents retained, for the discussion threads and the tweets, respectively. The proportions of stance-taking documents among the documents that were ranked by the algorithm as the top- $n$  documents most typical to the extracted topics are shown in Table 1. These were compared to the 95% confidence interval for the proportion of stance-taking documents among  $n$  documents randomly sampled from the corpus.<sup>1</sup> Measurements were carried out for  $n=35$  and  $n=100$ . The method used had yielded 90 re-run outputs, which each one of them contained a slightly different document ranking for the topics extracted. For each of the topics, we therefore extracted the 100 most top-ranked documents for every re-run, and ranked

<sup>1</sup>Calculated according to Preston (2012), i.e., 95% of all possible samples of size  $n$  yield a proportion  $\hat{p}$  within  $1.96\sqrt{p(1-p)/n}$  of the true proportion  $p$ .

Topics for discussion threads	Stance top $n$	
	$n=35$	$n=100$
people/think/mmr/like/child/really	71%	79%
rubella/women/immune/pregnant/girl	<b>94%</b>	<b>88%</b>
risk/child/disease/risks/carry/catching	82%	<b>90%</b>
immunity/herd/checked/wanes	85%	84%
mumps/meningitis/urabe/uk/mmr	88%	83%
000/10/offit/theory/cope/think/paul	74%	75%
children/damaged/unvaccinated	77%	83%
cough/whooping/brother/caught/mum	80%	80%
stance in entire corpus	80%	80%
corpus stance, 95% conf. interval	$\pm 13$	$\pm 7.9$

Topics for tweets	Stance top $n$	
	$n=35$	$n=100$
cancer/cervical/cause/prevention	42%	46%
girls/boys/10/need/vaccinated	<b>74%</b>	<b>73%</b>
vaccine/safe/child/effects/study	60%	<b>70%</b>
cancers/caused/related/prevent/protect	57%	45%
rhode/island/graders/mandates/7th	51%	51%
teens/getting/cdc/vaccinations	48%	<b>56%</b>
women/men/young/risk/infection/ask	48%	43%
vaccination/rates/low/states/adolescent	<b>74%</b>	<b>65%</b>
vaccinates/tdap/safety/teen	51%	<b>57%</b>
shot/got/doctor/tomorrow/arm	57%	52%
health/public/mandate/dept/activists	40%	41%
love/epidemic/documentary/television	14%	31%
vax/anti/age/cdc/proven/harm/govt	<b>80%</b>	<b>76%</b>
stance in entire corpus	44%	44%
corpus stance, 95% conf. interval	$\pm 16$	$\pm 10$

Table 1: The proportion of stance-taking documents among the top 35 and 100 most typical documents for each extracted topic. Each topic is represented by its most closely associated terms.

these documents according to the sum of the documents topic-association value over the 90 re-runs.

For the figures in Table 1, the stance proportion that lies below the 95% confidence interval for the stance proportions of  $n$  randomly selected documents is marked with italics and those that lie above are marked in boldface. That is, the document rankings (top 35 or top 100) that contain a smaller or larger density of stance-taking texts, than had the same number of documents been randomly selected, are marked in italics or boldface.

For discussion forum texts, for which the collection-level proportion of stance-taking was already high, the proportions among the documents extracted for the topics were similar to

the document-level proportion. The general trend was a slight increase in stance-taking documents, with one topic that had a stance-taking proportion above the 95% confidence interval for the top-35 documents and two topics that fulfil this criterion for the top-100 topics.

Also for the tweets, a majority of the topics had associated documents with a proportion of stance-taking that did not differ significantly from a random sampling from the document collection. However, some of the topics contained a very high proportion of stance-taking, in comparison to the proportion in the entire document collection. This resulted in that, for the tweets, there was a statistically significant difference for three topics also when extracting only the top 35 most typical documents. These top 35 documents were made up of document sets consisting of semantically coherent tweets. The topic *girls/boys/10...* mainly consisted of posts advocating HPV vaccine for both boys and girls, often also providing the argument that it prevents cancer. The documents belonging to *vax/anti/age/...* typically took the opposite stance, and often contained a questioning of whether there is a proof that HPV vaccination prevents cancer, or warnings against perceived adverse effects of HPV vaccination. The topic *vaccination/rates/low...*, which consisted of expressions of worries about HPV vaccination rates being low, forms an example of that stance-taking does not always imply that arguments are given. That is, although most of the tweets associated with this topic clearly take a stance in favour of vaccination, no direct arguments are given here.

There was also one tweet topic with a very low proportion of stance-taking among its associated documents, that is, the topic *love/epidemic/documentary...* which consisted of many tweets that, in different ways, but in a neutral manner, announced a documentary about HPV.

## 5 Discussion and conclusion

A typical practical application of the method studied here would be the case in which an analyst aims at finding frequently occurring vaccine-related arguments in a document collection that is too large for a fully manual analysis. The analyst would then instead perform an analysis at which only a subset of the texts would be read, i.e., those automatically extracted through topic modelling.

We have here shown, for the two document

collections investigated, that there are topics extracted which have associated documents that contain a larger proportion of vaccine stance-taking texts than the average document collection. That is, these document sets would form a useful resource for such an analyst who searches for vaccine-related argumentation.

The fact that there might also be topics extracted which do not contain argumentation, i.e., topics similar to the *love/epidemic...* topic, should not pose a large obstacle to the analysis, as long as there are other topics that have associated documents in which stance is taken. That is, at least for the documents extracted for the topic *love/epidemic...*, it is evident after reading only a few documents, that this topic is uninteresting for the task of finding argumentation. Documents closely associated with such topics can, therefore, be excluded from the analysis after a quick inspection. This would enable the analyst to focus on the other topics, which have associated documents that do contain argumentation.

## Acknowledgments

The study was funded by the Swedish Research Council (Vetenskapsrådet), through the project “Navigating in streams of opinions: Extracting and visualising arguments in opinionated texts” (No. 2016-06681).

## References

- Eric P. S. 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.
- Jingcheng Du, Jun Xu, Hsingyi Song, Xiangyu Liu, and Cui Tao. 2017. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *J Biomed Semantics*, 8(1):9.
- Kate Faasse, Casey J. Chatman, and Leslie R. Martin. 2016. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. *Vaccine*, 34(47):5808–5814.
- Lenny Grant, Bernice L. Hausman, Margaret Cashion, Nicholas Lucchesi, Kelsey Patel, and Jonathan Roberts. 2015. Vaccination persuasion online: A qualitative study of two provaccine and two vaccine-skeptical websites. *J Med Internet Res*, 17(5):e133.
- Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556 – 562.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Scott Preston. 2012. Math 158: Confidence interval for a proportion. <http://www.oswego.edu/~srp/158/CI%20Proportion/CI%20for%20a%20Proportion.pdf>.
- Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018. Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, number 247 in Studies in Health Technology and Informatics, pages 366–370. IOS Press.
- Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G Dunn. 2016. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *J Med Internet Res*, 18(8):e232.
- Timothy R. Tangherlini, Vwani Roychowdhury, Beth Glenn, Catherine M. Crespi, Roja Bandari, Akshay Wadia, Misagh Falahi, Ehsan Ebrahimzadeh, and Roshan Bastani. 2016. “Mommy blogs” and the vaccination exemption narrative: Results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR Public Health Surveill*, 2(2):e166.