Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-366

Vaccine Hesitancy in Discussion Forums: Computer-Assisted Argument Mining with Topic Models

Maria SKEPPSTEDT [a,b,1] Andreas KERREN [a] and Manfred STEDE [b]
[a] Computer Science Department, Linnaeus University, Växjö, Sweden
[b] Applied Computational Linguistics, University of Potsdam, Potsdam, Germany

Abstract. Arguments used when vaccination is debated on Internet discussion forums might give us valuable insights into reasons behind vaccine hesitancy. In this study, we applied automatic topic modelling on a collection of 943 discussion posts in which vaccine was debated, and six distinct discussion topics were detected by the algorithm. When manually coding the posts ranked as most typical for these six topics, a set of semantically coherent arguments were identified for each extracted topic. This indicates that topic modelling is a useful method for automatically identifying vaccine-related discussion topics and for identifying debate posts where these topics are discussed. This functionality could facilitate manual coding of salient arguments, and thereby form an important component in a system for computer-assisted coding of vaccine-related discussions.

Keywords. Vaccine hesitancy, topic modelling, argument mining

Introduction

Decreased vaccination rates caused by vaccine hesitancy have led to outbreaks of vaccine-preventable diseases in several parts of the world [1]. More might be learnt about the reasons for vaccine hesitancy by studying the arguments that are given for avoiding vaccination. The context of Internet discussion forums is one example of a context where vaccination-related arguments are expressed [2], and such forums are therefore one possible source from which vaccine-related arguments can be gathered.

There are a number of manual qualitative research methods that could be applied for coding Internet discussions on vaccination [3, pp. 163–180], and there are studies in which such a coding has been carried out [4, 5]. To be able to learn from and monitor Internet discussions on a larger scale, however, the content of large text collections needs to be coded. This is an intractable task when using manual coding approaches that require the entire text collection to be read, but is possible if important information could be automatically extracted and presented for manual coding. We here aim to explore this functionality, by performing computer-assisted extraction of arguments from Internet discussions on vaccination.[2]

[1] Corresponding Author: maria.skeppstedt@lnu.se.

[2] The study was funded by the Swedish Research Council, through the project "Navigating in streams of opinions: Extracting and visualising arguments in opinionated texts" (No. 2016-06681) and through the StaViCTA project, framework grant "the Digitized Society – Past, Present, and Future" (No. 2012-5659).

1. Background

For the task of coding a large text collection of free-text survey responses, Baumer et al. compared (i) an analysis based on a computer-assisted approach, where a subset of the texts was automatically selected by topic modelling and manually coded, and (ii) a manual analysis based on grounded theory, in which the entire document collection was coded [6]. The former analysis took a few hours to carry out by one analyst, while two researchers allocated several hours a week over about two and a half months for the latter approach. Despite the large difference in allocated time, the comparison of the output of the analyses showed that "The topic modeling results captured to a surprising degree many of the themes identified in grounded theory, and vice versa". These results show the potential of using topic modelling for selecting what material to manually code. We here followed the computer-assisted approach of Baumer et al., and manually coded a subset of our document collection, which we selected by topic modelling.

2. Method

Before the topic modelling algorithm was used to select documents to code, a pre-processing was applied to the texts in the collection.

2.1. The document collection used and the pre-processing applied

The texts that we used for exploring computer-assisted coding were vaccine-related discussion threads from the British parental website Mumsnet, which hosts online forums where subjects related to parenting are discussed.[3] The discussions are publicly available without a login, and debaters are encouraged to anonymise their texts, which makes it unlikely that the posts include sensitive or private content.

We have previously compiled a resource of debate posts from six Mumsnet discussion threads, where we have removed HTML-tags, names of debaters and citations from previous debaters [7]. We collected the debates with the criterion that they should have a title that indicates a debate topic related to vaccination or child vaccination in general, as opposed to more specific aspects of vaccination, e.g., vaccination against specific diseases. The publication year for the most recent post varied between the different threads, from year 2011 to 2017. In the previous study, we had also manually coded the posts as expressing a stance *for* or *against* vaccination, or as being *undecided*. For the current study, we were mainly interested in posts where an opinion was expressed. We therefore removed the undecided posts from the document collection, which led to a final collection of 943 posts. Each post was treated as one independent document in the experiment, and no meta-data was used. For instance, information on who the author was, or which thread the post belonged to, was not used.

The document collection was first automatically pre-processed by concatenating frequent collocations into one term. Different term instantiations of the same concept (morphological variations, synonyms, and related terms) were thereafter automatically replaced by a term that represented the concept. This was achieved by clustering [8] word embedding vectors that represented the terms. The embedding vectors were

[3] www.mumsnet.com/Talk

obtained from an out-of-the-box word2vec[4] model [9]. Table 1 shows examples of collocations (shown with an underscore) and concept clusters (shown with a slash). The produced concept clusters were manually corrected, which resulted in the removal of 165 terms from the clusters, and a final set of 402 concept clusters was used.

A standard stop word list [10] was used to remove stop words. We also extended this list by repeatedly pre-running the topic models to identify terms extracted by the model that we considered more suitable as stop words than as topic terms.

2.2. Topic modelling

The input to a topic modelling algorithm is typically a collection of text documents and the number of topics that the algorithm is expected to identify in the collection. The output of the algorithm, for each one of the identified topics, is (i) a set of terms from the collection that represents the topic, e.g., the terms in the first column in Table 1, and (ii) a ranking of the texts according to the probability that they discuss the topic.

The procedure of the previously mentioned topic modelling study by Baumer et al. [6] was followed. Baumer et al., however, only applied LDA (Latent Dirichlet Allocation) as the topic modelling method for their document collection of survey answers. Since previous research indicates that NMF (Non-Negative Matrix Factorisation) is more suitable to a document collection that consists of discussion posts [11], we also included NMF in the experiments. We instructed the topic modelling algorithms to identify ten topics, but only topics that were fairly stable over ten re-runs of the algorithm were retained. A 70% overlap of the returned term set with a previously returned term set was required for a topic to be considered stable. The experiments were implemented with Scikit-learn [10].

3. Results and discussion

The LDA algorithm produced very few stable topics. Only the output of the NMF algorithm, which returned six stable topics, was therefore analysed. For each of these topics, we extracted the 50 posts that the NMF algorithm ranked as most typical to the topic. One of the authors then manually coded the posts for arguments related to vaccination. A few hours were spent on each topic, and the results are shown in Table 1.

A set of semantically coherent arguments could be identified for each extracted topic. Topic 2 was the most coherent of the six topics, as only 23 themes were identified, which all of them were related to Dr. Paul Offit. Both Topic 1 and Topic 4 were related to MMR (measles, mumps, and rubella) vaccination, but the themes of Topic 1 were related to research on and reports of adverse vaccine reactions, while the Topic 4 posts discussed the duration of vaccine immunity, disease severity, and single vaccines. Topic 3 was related to the eradication of diseases through vaccinations, opinions on how small pox was eradicated and how that should affect vaccination programs for other diseases. The arguments for Topic 6 revolved around risk assessments for child vaccination, for vaccine-preventable diseases, and for infecting vulnerable individuals.

[4] https://code.google.com/archive/p/word2vec

Table 1. Arguments occurring in at least three posts in the layperson discussions studied (without any regard for their validity). For the 6 topics, 38, 23, 51, 61, 40, and 33 themes, respectively, were identified.

Extracted terms	Arguments (number of occurrences)
Topic 1: mmr, evidence, link, worry/concern, problem/difficulty, autism/autistic, parents, thread, sure, case, jab, issue, study, reaction/ response, research, pro, opinion/views, wrong, reason, information	There is no proven link between MMR and autism, despite many studies (7). Children who have been reported as having reacted badly to MMR are not examined/parents dismissed/more research needed (6). Many reports of vaccination damage and regressive autism after vaccination (4). MMR effective in stopping the spread of 3 potentially dangerous diseases/Children vulnerable without (3). Personal story of measles after vaccination (3). Personal story of immune system negatively affected by MMR/measles vaccination (3). The scientific papers that suggest a link between MMR and autism have been shown invalid (3). Expression of distrust in government/pharmaceutical industry (3).
Topic 2: offit, 10, 000_ vaccines, theory, baby/ infant, antigen/epitope, cope, 000, paul, agree, book, goon, test, bullshit/nonsense, 100, theoretical, flawed, day	Criticism of Offit's claim that each infant would have the theoretical capacity to respond to about 10,000 vaccines at any one time (23). Explanation/defence of Offit's claim about the theoretical response to 10,000 vaccines (8). Offit is lobbying for the pharmaceutical industry (5). Offit is a good person (5). Offit makes money on vaccines, and therefore biased in vaccination debates (3). Request for evidence of risks with vaccination combinations, rather than criticism of Offit (3).
Topic 3: small_pox, countries, eradicate, mass_vaccination, early/late, anybody/ somebody, endemic, population, polio, complex, epidemic/ outbreak, current, poor	That small pox vaccination has been successful does not mean that there are no problems with other vaccines (7). Small pox was eradicated by the vaccine (6). No proof that vaccination was responsible for eradicating small pox, but it could have been caused by other factors, e.g., better sanitary and health care conditions (6). Better conditions is not the reason small pox was eradicated, since it was eradicated also in poor countries without these improvements (6). We are close to eradicating polio through vaccination (5). Selective vaccination was enough to eradicate small pox, no need for mass vaccination to eradicate a disease (3).
Topic 4: mumps, dose, measles, wanes, waning, protect, mumps_ vaccine, complication, caught, cases, student, introduce, groups, meningitis, difficulty/ problem, single, age, mmr, singles, economic	Duration of protection from mumps vaccine is unknown or uncertain (9). Not only the MMR combination should be offered, but also single vaccinations (5). Mumps is more dangerous in adulthood (4). Infant mumps vaccination increases risk for outbreaks of mumps in older age groups (4). Mumps vaccination is motivated by economics, not health reasons (4). There are serious complications from mumps, e.g., sterility, meningitis and deafness (4). Mumps in children is a mild disease with few complications (3). Vaccination decreases risks for complications, even if immunity wanes (3). Complications caused by mumps are rare (3).
Topic 5: doctor/ physicians, assumption, anti, smoking, digging, single, critical, parents, patient, swine_flu, fringe, refusal, medical_ professionals, article, switzerland, loibner, austria, shows, lie	Expression of trust in science and medical professionals or a criticism of distrust in science and medical professionals (5). Expression of distrust in medical professionals (4). Expression of distrust in the pharmaceutical industry, e.g., unethical, biased in information given (4). There are physicians who do not vaccinate themselves or let their children be vaccinated (4). Many reports of vaccination damage and regressive autism after vaccination (3). A questioning of the claim that there is a recent trend of vaccine hesitancy among physicians (3). There are other ways than vaccination to protect others who are vulnerable, e.g., quarantine (3).
Topic 6: risk, benefit, small/tiny, end, carry_ risk, vulnerable/prone, surgery, effects, higher, case, worth, catch, parent, real, protection, rare, healthy_child, decision, accept, choose/ decided, minimal, community, disease/ infectious_diseases, im- moral, herd_immunity	The risk of catching the vaccine-preventable disease or that the disease will result in complications is higher than the risk of the vaccination (17). Vaccination can protect others who are vulnerable/contribute to herd immunity (10). The risk of vaccination is higher than the risk of the disease (6). There is no need to take the risk of vaccination, since vaccination is unnecessary (6). A child should not be vaccinated with the aim of protecting others (4). The children that are vulnerable to vaccination damages cannot be identified beforehand/no screening done to identify them, and for those children vaccination carry a high risk (4). Parents are not appropriately informed about risks of vaccination (3). The risk of serious side effects from vaccination for an otherwise healthy individual is minimal (3). Parents' primary responsibility is their own child, not to protect others (3).

Topic 5 was related to trust or distrust in the medical profession and industry, and to attitudes towards vaccination within the medical profession. This topic was, however, less semantically coherent than the others, and included many examples of themes not related to these issues. It can also be noted that both arguments *for* and *against* vaccination were identified, for all six topics analysed.

4. Conclusion and future directions

The semantic coherence of the texts analysed for each topic, and the fact that reoccurring arguments were found among these texts, indicate that the application of NMF-based topic modelling is a useful strategy for extracting frequently occurring discussion topics and salient arguments. As these topics and arguments were extracted by only analysing a subset of the text collection, this method is suitable for computer-assisted analyses of Internet discussions on a larger scale, with the potential of finding reasons upon which vaccine hesitancy is based.

The study does, however, not show that the topics and arguments extracted were the *most* salient ones. Future work will therefore include a manual annotation of randomly selected texts to determine if there were important topics not included in our analysis. It should also be noted that programming skills were required to perform the pre-processing and topic modelling. As this limits the usability of the approach studied, we aim to develop an interactive, graphical tool by which computer-assisted argument mining with topic models can be carried out.

References

[1] Larson HJ, Ghinai I: Lessons from polio eradication. Nature 2011, 473(7348):446–7.
[2] Campbell H, Edwards A, Letley L, Bedford H, Ramsay M, Yarwood J: Changing attitudes to childhood immunisation in English parents. Vaccine 2017, 35(22):2979–2985.
[3] Myers MD: Qualitative research in business & management. London: SAGE 2009.
[4] Skea ZC, Entwistle VA, Watt I, Russell E: 'Avoiding harm to others' considerations in relation to parental measles, mumps and rubella (MMR) vaccination discussions — an analysis of an online chat forum. Soc Sci Med 2008, 67(9):1382–90.
[5] Faasse K, Chatman CJ, Martin LR: A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016, 34(47):5808–5814.
[6] Baumer EPS, Mimno D, Guha S, Quan E, Gay GK: Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? J Assoc Inf Sci Technol 2017, 68(6):1397–1410.
[7] Skeppstedt M, Kerren A, Stede M: 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), Stroudsburg, PA, USA: Association for Computational Linguistics 2017:1–8.
[8] Ester M, Kriegel HP, Sander J, Xu X: A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Palo Alto, California, USA: AAAI Press 1996:226–231.
[9] Rehurek R, Sojka P: Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Paris, France: European Language Resources Association (ELRA) 2010:45–50.
[10] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et.al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011, 12:2825–2830.
[11] Sobhani P, Inkpen D, Matwin S: From Argumentation Mining to Stance Classification. In Proceedings of the 2nd Workshop on Argumentation Mining, Stroudsburg, PA, USA: Association for Computational Linguistics 2015.