MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190241

Finding Reasons for Vaccination Hesitancy: Evaluating Semi-Automatic Coding of Internet Discussion Forums

Maria Skeppstedt[a,b], Andreas Kerren[a], Manfred Stede[b]

[a] *Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden,*
[b] *Applied Computational Linguistics, University of Potsdam, Potsdam, Germany*

Abstract

Computer-assisted text coding can facilitate the analysis of large text collections. To evaluate the functionality of providing an analyst with a ranked list of suggestions for suitable text codes, we used a data set of discussion posts, which had been manually coded for reasons given for taking a stance on the topic of vaccination. We trained a logistic regression classifier to rank these reasons according to the probability that they would be present in the post. The approach was evaluated for its ability to include the expected reasons among the n top-ranked reasons, using an n between 1 and 6. The logistic regression-based ranking was more effective than the baseline, which ranked reasons according to their frequency in the training data. Providing such a list of possible codes, ranked by logistic regression, could therefore be a useful feature in a tool for text coding.

Keywords:
Vaccination Refusal, Text Mining, Supervised Machine Learning

Introduction

Vaccination hesitancy has led to outbreaks of vaccine-preventable diseases in several parts of the world [1]. To study reasons that are given for vaccination refusal and vaccination hesitancy in different types of user-generated text resources, e.g., Internet discussion forums, might be one method for increasing our knowledge about this phenomenon.

There are previous studies in which these types of texts have been manually analysed for reasons for vaccination hesitancy [2, 3]. There are also studies which have applied topic modelling for automatically extracting vaccination-related information from text collections too large for a fully manual analysis [4, 5]. Previous research has shown topic modelling to be an efficient text-mining method for selecting and topically sorting texts, but a manual coding of the selected texts is also recommended for a deeper understanding of their content [6].

We have, in a previous study, performed such a manual coding of texts selected through topic modelling. We had applied topic modelling on Internet discussions of the subject of vaccination, and the topic modelling algorithm automatically extracted six topics present in the collection. We then manually coded 50 texts associated with each one of these topics, through identifying the reasons that the authors had given for taking a stance *for* or *against* vaccination. We detected 242 different unique reasons in total, some only occurring once and some reoccurring [7]. Although the text selection and sorting that was provided by the topic modelling facilitated the coding, it was still a demanding task. It was particularly difficult to determine whether a reason identified in the text was a new one, not yet included in the analysis, or whether this reason had occurred in any of the previously analysed texts.

The coding procedure might, therefore, be simplified if the user is assisted in the task of locating suitable text codes among the codes previously identified. That is, when an analyst selects a text for a manual coding of reasons, a system could automatically suggest which previously identified reasons that might be found in the text that is being coded. The system could, for instance, rank previously identified reasons according to how likely it is that they are present in the text. Such a ranked list could support the analyst in the task of determining whether a reason identified in the text that is being coded has occurred in any of the previously analysed texts.

The aim of the study described in this paper is to design and evaluate a method for providing such a ranked list of previously identified reasons.

Previous Studies on Reason Classification and Identification

Among the large body of recent research on stance detection and argumentation mining, the ones most relevant for the present study investigate a task that is referred to as argument identification or recognition [8, 9] or as reason identification and classification [10]. These studies have either (i) analysed a text collection for all reasons occurring in the collection, and then not included the infrequently occurring ones in the automatic classification experiments [10], or (ii) used external resources for finding a few important reasons that are likely to occur in the text collection, and then labelled the texts with these reason categories [8, 11]. When training classifiers on four different text collections to recognise 11, 12, 15, and 18 reason categories, respectively, F-scores in the low 50s were achieved [10]. When instead training classifiers on two different text collections to recognise six and seven reason categories, respectively, micro-average F-scores between 0.7-0.8 were achieved [8]. Topic-independent classifiers were created by using training data that was constructed through measuring the similarity between the opinionated text and the text describing the reason. Finally, an F-score of 0.49 was achieved for the task of classifying texts into 16 reason categories, by using the method of clustering texts according to 16 topics derived from topic modelling based on Non-negative Matrix Factorization (NMF) [11]. The automatically

extracted topics were manually mapped to 16 pre-defined, prominent reasons. There were also previous attempts to identify prominent reasons in texts in a fully unsupervised fashion, which however have yielded low results [9].

In contrast, for the use case studied here, also the less frequently used reasons must be included. A final report of the outcome of the text coding might only include frequently occurring reasons; however, during the process in which the analyst carries out the text coding through identifying reasons in the texts, it is still unknown which reasons that occur frequently in the text collection. Both the prominent and non-prominent ones among previously identified reasons are therefore relevant to identify in the text that is in the process of analysis. Consequently, both prominent and non-prominent reasons should be provided as coding suggestions to the user, and should therefore also be included in the experiment.

Methods

The experiments consisted of training a classifier to rank the reasons that had previously been identified in the coding, according to the likelihood that they are present in the text that is currently coded.

The Previously Constructed Data Set

Our previous study consisted of applying the NMF topic modelling algorithm on posts from vaccine-related Internet discussions from the British parental website Mumsnet [7]. We chose the data based on the fact that the discussions were taking place on a general discussion forum for parents. Although views expressed in the Mumsnet forum by no means could be claimed to be representative for a population other than its actively participating users, the opinions expressed on Mumsnet are likely to be more general than, e.g., opinions expressed on an anti-vaccination website. In addition, Mumsnet informs their users of the fact that the site's discussion threads are published as publicly available posts, for which no login is required. Debaters are also asked to anonymise their postings, e.g., by using a chat nickname. Therefore, according to the data classifications provided by Eysenbach and Till [12], the content of the forum belongs to a public sphere rather than a private one.

Before applying topic modelling, we pre-processed the texts by removing stop words and performing a concept cluster-detection. That is, we concatenated into one term word collocations that occurred frequently in the text collection, and we replaced different term instantiations of the same concept,

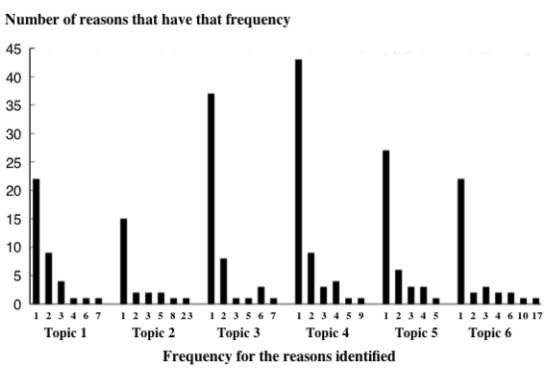

Figure 1 – The distribution of the frequency of the unique reasons identified for the six topics.

e.g., synonyms and different inflections of a word, with a common string representing this concept. The term-replacement was achieved by the use of word embedding vectors associated with the terms in the corpus. These vectors were clustered using DBSCAN clustering [13], and terms whose vectors were assigned to the same cluster were considered as belonging to the same concept. We chose DBSCAN, since it uses the properties of the data set that is to be clustered to determine how many clusters to create, and the user thereby does not have to provide an a priori specification of the number of appropriate clusters. A clustering algorithm with this property is a requirement for this clustering task, as the number of clusters of inflections and synonyms that occur in the text collection is unknown beforehand. We used a final cluster set of 402 clusters, which we obtained by a manual quality control of the automatically constructed clusters. Through the quality control, 165 terms were added to a list of terms to exclude in the cluster construction process.[1]

The topic modelling algorithm automatically extracted six topics from the text collection. The subject of Topic 1 and 4 was Measles, Mumps, and Rubella (MMR) vaccination; for Topic 1, it was in the form of reasons related to fear of and research on adverse vaccine reactions, and for Topic 4, it related to vaccine immunity duration, disease severity, and single vaccines. Topic 2 was about the vaccinology professor Paul Offit, Topic 3 was about eradication of diseases through vaccinations, and Topic 5 contained reasons related to trust/distrust in the medical profession and industry, as well as reasons related to attitudes towards vaccination among medical professionals. Finally, texts for Topic 6 discussed risk assessments for vaccination, for diseases, and for infecting others.

For each one of the topics extracted by the topic modelling algorithm, we manually coded 50 texts for reasons that were given *for* and *against* vaccination. This analysis consisted of reading the selected texts and identifying reasons mentioned in the text. A reason identified could either be (i) a reason that had not previously occurred in the text collection, or (ii) a reason previously identified in the text collection. In the first case, we wrote a new description of this reason in the coding sheet, and in the second case, we noted the exact same description as previously used in the coding sheet. We performed the analysis in Microsoft Excel, with one column for the texts and additional columns for the descriptions of the reasons identified in the texts.

Most of the coded descriptions of reasons were formulated as an argumentative statement, e.g., "*There is no proven link between MMR and autism, despite many studies*", "*That small pox vaccination has been successful does not mean that there are no problems with other vaccines*", "*Not only the MMR combination should be offered, but also single vaccinations*", and "*The risk of catching the vaccine-preventable disease or that the disease will result in complications is higher than the risk of the vaccination*". Some were, however, formulated as a meta-description of reasons, e.g., "*Expression of distrust in government/pharmaceutical industry*". The original study reported on the detailed results of the coding [7].

The number of reasons identified varied between the different topics, as did the frequency of occurrence for the reasons (Figure 1). For a few texts, no reasons were detected, and these were removed from the data set used in the experiments for the present study. This resulted in a final data set for the six topics that contained 42, 47, 50, 49, 49 and 50 texts,

[1] Examples of word pairs clustered together are: worry/concern, problem/difficulty and autism/autistic

respectively, and where each one of them had one or several associated reasons.

Classifier Training

We carried out six separate experiments for each one of the extracted topics. That is, classifiers were trained and evaluated on texts and reasons that belonged to one topic at a time. In addition, we also performed an experiment on a data set in which the six topic-divided data sets were combined into one set. This evaluation used leave-one-out cross-fold validation.

The type of machine learning classifier chosen for the experiment was a logistic regression classifier. This choice was based on the inherent data scarcity of the task evaluated. That is, the task of assisting a text coder in locating the appropriate reason category among a set of previously identified reasons required a classifier suitable for small training data sets. This inherent data scarcity is due to that, in the typical use case, most of these previously indentified reasons will have a limited number of associated texts that can be used for forming the training data set.

Moreover, the logistic regression classifier returns a probability estimate for each one of the possible classification categories in the data set. This probability estimate is not only suitable to use for ranking the previously created reasons that the coding process suggests to the user, but it also forms an output that is human interpretable. The probability estimate of a logistic regression classifier could therefore be shown to the user as an indication of the likelihood that a previously identified reason is present in the text that is currently being coded. This would not be the case for classification outputs such as the distance to the separating hyperplane of a support vector machine, which would make little sense to a human text coder.

The situation simulated in the experiments was that all texts associated with a topic would have been coded for reasons, except one of them. That is, a situation in which the left out data point in the leave-one-out scheme would not yet have been coded. We used the rest of the data, i.e., all data points except the one left out, for training a logistic regression classifier. Features were extracted from the texts, and their associated reasons were used as classification categories. Texts with several associated reasons were added multiple times to the training data set, once for each of its associated reasons, with that reason as the classification category. The describing text for each one of the coded reasons was also added to the training data, to form one additional training data point for each one of the classification categories.

Tokens that occurred at least twice in the data set were used as features, and the same stop word list as had been used for constructing the topic models was applied to remove stop words. Due to the scarcity of data, no n-grams were used. The probability estimate for each classification category that is returned by the logistic regression classifier was used for producing a ranked list of reasons, which thus formed the output of the classification. That is, the larger the probability that a text would be classified as containing a certain reason, the higher ranking did this reason achieve in the ranked list produced for this text.

Apart from the standard method of using tokens as features, we also performed an experiment with the same term-cluster replacement as used for the topic modelling. That is, different terms that belong to the same concept were replaced with a unique string representing this concept. This concept-string was then treated as a normal token by the feature extraction procedure.

The logistic regression classifier available in scikit-learn was used with default parameters [14], as we considered the data available too small for parameter tuning. However, we ran all experiments with the default inverse L2 regularisation strength of 1 as well as with a strength of 10. We used the Gensim library [15] for accessing embedding vectors and used an out-of-the-box word2vec model,[2] trained on Google news.

Evaluation

Results for the logistic regression classifier with standard features, as well as for the same classifier with cluster features, were recorded for both L2 settings. We applied leave-one-out cross-fold validation on all evaluations.

The situation that the analyst needs to find a reason among previously identified reasons only arises when a reason contained in the text already has been identified in at least one of the previously analysed texts, i.e., not when a text is going to be assigned to a *new* reason. For reasons that have exactly one text association in the data set, the only situation that arises is this one where a new reason is created. A ranking of existing reasons thereby does not make any sense for this situation. Therefore, this evaluation of the reason ranking did not include these text-reason associations as left out data points. The associations were however included in the training data, and thereby as possible classification categories. This is in accordance with the authentic situation, in which it would be unknown to the classifier that the reasons that only occur once in the data are not to be associated with the held-out data point. When performing the classification for the left out data point, these reasons are therefore just as valid as any of the other reasons to include in the ranking. The inclusion of these data points makes the evaluation more realistic, as the task would have been simplified if these reasons had been removed altogether from the experiment.

The ranking was evaluated on the criterion of the proportion of cross-validation folds for which the expected reasons were among the *n* top-ranked reasons. An *n* ranging from 1 to 6 was used, as up to around six reasons would be a reasonable number of coding suggestions to scan through when coding a text.[3]

Baseline and Ceiling

The aim of the study was to find out whether the use of machine learning would be an appropriate approach for solving the task of providing the user with coding suggestions. We therefore focused the study on evaluating the performance of the machine learning classifier we deemed most suitable for this task, rather than on comparing different machine learning algorithms. We compared this classifier, i.e., a logistic regression classifier, to a baseline strategy that did not employ machine learning. This baseline strategy instead ranked the reasons according to their frequency, i.e., according to the number of texts they were assigned to in the training data.

Note that the baseline strategy ranked the texts according to the reason frequency in the training data set, not the entire data set. This means that the reason frequencies that the held-out data point contributed to were not a part of the frequency count, as these would not be known in a realistic setting. This, in turn, results in that the baseline rankings sometimes differ depending on which data point is used as held-out data. It is

[2] code.google.com/archive/p/word2vec/ (March 27 2019)

[3] The experiments can be replicated by running the Python script *run_classifier.py*, found at https://github.com/mariask2/topics2themes (March 27 2019).

therefore possible for the baseline to, for each evaluation fold, fail to produce a correct ranking.

Since the same text often is associated with several reasons, it is not possible to achieve a score of 100 percent correct inclusion of reasons when the evaluation criterion only takes the most top-ranked elements into account. That is, not all *a* reasons that are associated with a held-out text can be included among the *n* top-ranked elements for the cases when $a > n$. For each of the six topics and each of the top *n* rankings for which results were evaluated, we therefore calculated an upper ceiling. If *n=1*, and there is only one associated reason (*a*) that belongs to the text (*a=1*), an optimal ranking will give that reason the highest-ranking position and contribute to one correct classification. However, if there, e.g., are three reasons associated with a text (*a=3*), two of these reasons can *not* be ranked highest, and an optimal ranking of reasons for that text will therefore contribute with one correct classification and two incorrect ones, or generally: for each text, if $a > n$, the number of incorrect classifications of an optimal ranking are equal to $a - n$.

Results

Table 1 below shows results for each one of the six topics as well as for the combined data set. The results show the proportion of times that the expected reason was included among the *n* top-ranked reasons (where *n* ranges from 1 to 6). Four numbers are presented, which includes the standard and cluster features for logistic regression (for an inverse L2 of 10) and the frequency-based baseline ranking and ceiling results, for each one of the points of measure. The boldfaced font indicates the best result among the three ranking methods.

Both of the logistic regression-based methods performed better than the baseline method, except for three points of measure (shown by underscore in Table 1). For one of these points of measure, the lower-performing of the two logistic regression-based methods performed equally to the baseline method, and for the other two points of measure, the lower-performing of the two logistic regression-based methods achieved a result that was two percentage points lower than the result achieved by the baseline. Despite these few exceptions, the logistic regression-based methods clearly outperformed the baseline method with general and large performance differences.

The difference was less evident between standard logistic regression and logistic regression with concept clusters. Which of the two methods performed best varied between different topics and between different values of *n*. Also the regularisation strength had a very limited effect on the overall results. For half of the topics, the results decreased a few percentage points with a stronger regularisation, while it led to a minor increase in results for the other half of the topics.

Discussion

There is a large variation among the six topics in the data set. That is, a variation (i) in terms of the number of identified reasons (from 23 to 61), (ii) in the frequency distribution of these reasons (see Figure 1), (iii) to what extent several reasons were associated with the same text (as indicated by the ceiling calculations), and (iv) in terms of what classification results were achieved. That the logistic regression methods generally performed better than the frequency-based ranking for this diverse data set indicates some level of generalisability of the results, despite the limitation that the experiment was performed on one single data set. Although the actual efficiency of the ranking might differ between different text collections and different manual coding strategies, the results achieved encourage us to implement text coding support by using logistic regression-based ranking of previously identified reasons.

In contrast, since the use of concept-cluster features does not yield any obvious advantage, there is no point in carrying out the more complex feature extraction process that the use of these features entail.

It should be noted that the situation evaluated here is when all but one of the texts already have been assigned reasons, and that other results might have been achieved if an earlier point in the analysis process had been simulated. To perform such a simulation is a task to include in future work. It might also be interesting to, similar to previous work [8], evaluate the performance of a ranking that relies entirely on the analyst's description of the identified reason; however, such an approach might be less suitable for detecting reasons that are formulated in the form of meta-descriptions.

We have previously constructed an interactive visualisation tool whose graphical user interface displays the output of topic modelling applied to a text collection [16]. The tool displays the texts selected by the topic modelling algorithm, and includes methods for searching and sorting among these texts. The tool also provides the functionality of attaching user-defined *themes* to the texts. That is, for the application described here, these themes would correspond to reasons identified when analysing texts. The interactive tool does, however, not yet include any functionality for suggesting which themes to associate to a text that is being analysed. Instead, the tool shows themes in a list previously identified by the user, which is simply sorted according to the creation time of the themes. The next step in developing the tool will therefore be to add the text analysis support investigated here.

*Table 1 – The proportion (%) of evaluated texts for which the expected arguments was found among the n top-ranked reasons (**log**: a logistic regression classifier with standard features, **clu**: a logistic regression classifier with cluster features, and **bas/cei**:baseline/upper ceiling).*

% found	Topic 1 (38 reasonss)			Topic 2 (23 reasons)			Topic 3 (51 reasons)			Topic 4 (61 reasons)			Topic 5 (40 reasons)			Topic 6 (33 reasons)			All topics (242 reasons)		
Rank	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*	*log*	*clu*	*bas/cei*
n=1	**28**	21	15/61	**53**	**53**	45/71	35	**39**	0/58	32	**33**	16/49	**34**	**34**	0/74	**32**	28	28/61	29	**30**	8/61
n=2	**34**	32	28/93	**65**	**65**	61/91	45	**49**	14/81	**47**	46	16/76	**50**	47	13/91	43	**48**	45/89	**41**	**41**	13/86
n=3	**45**	34	28/100	**76**	**76**	61/100	57	**61**	14/90	**63**	60	16/88	**58**	55	16/97	**63**	53	50/98	**54**	51	16/95
n=4	**49**	40	30/100	**80**	78	**80**/100	67	**69**	22/95	**74**	70	28/95	63	**66**	24/100	**75**	70	65/100	**61**	57	16/98
n=5	**51**	**51**	36/100	**84**	82	80/100	**73**	71	61/98	**77**	74	30/100	**71**	68	24/100	80	**82**	68/100	**64**	62	16/100
n=6	**57**	**57**	36/100	**88**	**88**	80/100	**82**	76	61/99	**79**	**79**	44/100	**74**	68	26/100	**83**	82	72/100	**68**	66	20/100

That is, to use a logistic regression classifier to rank the previously identified themes in the list according to the likelihood that they are present in the text that is currently being analysed. Such a classifier could be re-trained, whenever the user creates a new text-theme association, and thereby be able to continuously adapt and improve the theme ranking when the user extends and refines the manual analysis.

We hope that such a tool, with the extended coding support investigated here, could be useful for assisting coding of recurring content in large text collections. For instance, for coding discussion forums for recurring arguments used in vaccination debates. The semi-automatic approach of such a tool, i.e., to extract texts typical to frequently occurring topics and then assist the user in the task of manually coding the extracted texts, does not guarantee that all important subjects discussed in the text collection are detected. However, the approach offers the possibility of accessing important content in text collections that are too large to make a fully manual coding feasible. The output of the semi-automatic coding could, for instance, be used for forming hypotheses on reasons why different types of health decisions are made, e.g., health decisions related to vaccination.

Conclusions

We have investigated the efficiency of ranking previously identified reasons according to the likelihood that they are present in a text that is being manually coded. The performance of a logistic regression classifier that ranked reasons according to the probability that they would be present in the text that is being coded was compared to a ranking based on the frequency of the reasons in the data set. When evaluating the approaches for their ability to include the expected reason among *n* top-ranked reasons, using an *n* between 1 and 6, we could conclude that the logistic regression-based ranking outperformed the frequency-based one. These results therefore encourage us to construct a system that provides the human text coder with a ranked list of suggestions for possible reasons that might be present in a text, and to implement this ranking by the means of a logistic regression-based classifier.

Acknowledgements

We would like to thank the Swedish Research Council (Vetenskapsrådet) that funded this study through the project "Navigating in streams of opinions: Extracting and visualising arguments in opinionated texts" (No. 2016-06681). We are also very grateful for the useful input provided by the reviewers.

References

[1] H.J. Larson, I. Ghinai, Lessons from polio eradication, *Nature* **473** (7348) (2011), 446–7.

[2] L. Grant, B.L. Hausman, M. Cashion, N. Lucchesi, K. Patel, J. Roberts, Vaccination persuasion online: A qualitative study of two provaccine and two vaccine-skeptical websites, *J Med Internet Res* **17**(5) (2015), 133.

[3] K. Faasse, C.J. Chatman, L.R. Martin, A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post, *Vaccine* **34**(47) (2016), 5808–5814.

[4] T.R. Tangherlini, V. Roychowdhury, B. Glenn, C.M. Crespi, R. Bandari, A. Wadia, M. Falahi, E. Ebrahimzadeh, R. Bastani, "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) (2016), 166.

[5] D. Surian, D.Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, A.G. Dunn, Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection, *J Med Internet Res* **18**(8) (2016), 232.

[6] E.P.S. Baumer, D. Mimno, S. Guha, E. Quan, G.K. Gay, Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?, *Journal of the Association for Information Science and Technology* **68**(6) (2017), 1397–1410.

[7] M. Skeppstedt, A. Kerren, M. Stede, *Vaccine Hesitancy in Discussion Forums: Computer-Assisted Argument Mining with Topic Models*. In Proceedings of Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth, Studies in Health Technology and Informatics, IOS Press, (2018), 366–370.

[8] F. Boltužic and J. Šnajder, *Back up your Stance: Recognizing Arguments in Online Discussions*. In Proceedings of the First Workshop on Argumentation Mining, Association for Computational Linguistics, Stroudsburg, PA, USA, (2014), 49–58.

[9] F. Boltužic, J. Šnajder, *Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity*. In Proceedings of the 2nd Workshop on Argumentation Mining, Association for Computational Linguistics, Stroudsburg, PA, USA, (2015), 110–115.

[10] K.S. Hasan, V. Ng, *Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, (2014), 751–762.

[11] P. Sobhani, D. Inkpen, S. Matwin, *From Argumentation Mining to Stance Classification*. In Proceedings of the 2nd Workshop on Argumentation Mining, Association for Computational Linguistics, Stroudsburg, PA, USA, (2015).

[12] G. Eysenbach, J.E. Till, Ethical issues in qualitative research on internet communities, *BMJ (Clinical research ed.)*, **323**(7321) (2001), 1103-5.

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *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, AAAI Press, Palo Alto, California, USA, (1996), 226–231.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.

[15] R. Rehurek, P. Sojka, *Software Framework for Topic Modelling with Large Corpora*. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, European Language Resources Association (ELRA), Paris, France, (2010), 45–50.

[16] M. Skeppstedt, K. Kucher, M. Stede, A. Kerren, *Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics*. In Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources, European Language Resources Association (ELRA), Paris, France, (2018), 9–16.

Address for correspondence

Maria Skeppstedt: maria.skeppstedt@lnu.se.