Veronika Laippala, Aki-Juhani Kyröläinen, Aleksi Vesanto, Lotta Lehti, Johanna Kallio, Filip Ginter. University of Turku

Welfare bums or fellow humans: Discussing welfare provision to the poor in Suomi24 forum

We present an ongoing study which profits from the Suomi24 data resource. The study is part of the project *Whose welfare state?* at the University of Turku. The project combines discourse analysis, sociology and natural language processing to explore people's expression of attitudes towards the welfare state system in different contexts ranging from institutional documents and media texts to internet discussion fora. In sociology and social policy, attitudes towards different dimensions of the welfare state are mostly examined through survey data (see, however, Saari, Behm & Lagus 2017). Our project brings new perspectives to the study of these attitudes by the use of naturally occurring data.

In this paper, we focus on exploring how the general public discusses welfare provision to the poor. According to previous studies on survey data, Finns overall do not blame the poor for their situation and consider supporting them as legitimate (e.g., Niemelä 2008). However, Laihiala & Ohisalo (2017) show that online discussions feature a different attitude: blaming the poor and delegitimation of social allowance is common. Similar findings have been reported by Saari & al. (2017) and Suur-Askola (2017).

In order to profit from the entire 2,3 billion token Suomi24 corpus, we applied topic modeling, an unsupervised machine learning method to explore large volumes of unlabeled text (e.g., Blei, Ng, & Jordan 2003; Rehurek & Sojka 2010). Topic modeling is widely applied in many fields utilizing large language resources, such as digital discourse analysis, social and political sciences and journalism. For instance, topic modeling has been shown to be helpful in identifying important news items (Krestel & Mehta 2010) and in exploring the development of news article topics over time (Jacobi et al. 2013). In this study, we used structural topic modeling (STM), implemented in R (package stm, version 1.3.0). In this model, a topic is specified as a mixture over words and each word is associated with a probability of belonging to a topic. Similarly, a document constitutes a mixture of topics, i.e., a given document can consist of multiple topics. Additionally, STM allows the inclusion of metadata associated with the documents. For instance, time can be included as a covariate that can either influence the prevalence or the content of the topics (see Roberts et al. 2013; Roberts et al. 2016a).

In order to set up the data for the study, we first extracted from the Suomi24 corpus all discussion comments that included the lemma *köyhä* 'poor' or one of its 14 near synonyms. The lemmas were retrieved with a version of the data set analyzed with the Finnish Dependency Parser (Luotolahti et al. 2015) and the near synonyms were identified with the Word2Vec algorithm (Mikolov et al. 2013). Finnish models are available at http://bionlp-www.utu.fi/wv_demo/. This yielded us a corpus of 378,387 comments. In this talk, we focus on the comments published in 2014. After relatively heavy preprocessing to clean the data from duplicates and linguistically uninteresting material such as punctuation and pronouns, the final data set consisted of 32,407 comments.

Then, STM was fitted to the final data set. We also included two covariates to the model: length of the comment to capture linguistic complexity and the month of the posting to capture potential seasonal differences. To estimate the number of topics for these data, we used a spectral initialization method which has been shown to offer good performance on large data sets (see Roberts et al. 2016b). A solution with 46 Topics was estimated to have the best fit to the data.

In the exploration of the Topics, we firstly analyzed 25 keywords estimated for each Topic. This revealed that the topic modelling solution makes sense: the keywords create semantically meaningful groupings which reflect different areas of public debate concerning welfare provision to the poor. Among others, these areas include political decisions, religious principles and consumption habits.

Second, we examined the co-variation of the length of the comment with the Topics. Preliminary analysis of some of the Topics give promising indications on the relation of the topic with the comments' pragmatic functions. For example, the prevalence of the Topic number 28 (vulgar comments judging the poor) was negatively correlated with the length of the comment while the effect of length was not statistically significant with the Topic number 36 (sharing news concerning food aid lines). Shorter comments thus seem to be related with insults and strong judgement. Further, both topics were strongly influenced by the month of the posting: the prevalence of the topics degreased after January but started to rise again during the summer and peaked between August and September. This seasonal effect was far stronger for Topic 36.

In order to understand more closely how the Topic model solution reflects the debate on welfare provision, we will compare the Topics with the sociological criteria of deservingness (see Van Oorschot 2006): control over neediness, level of need, identity, attitude, and reciprocity. These criteria are often used to describe the deservingness of help of different social groups: who deserves to be helped and why. We will analyse which criteria can be identified in the different Topics and in which manner. Overall, our objective is to provide social scientists and policy makers with useful information concerning attitudes towards welfare provision as expressed in the Suomi24 discussion forum.

References

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

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. 2013. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1) 2013.

Krestel, Ralf, and Bhaskar Mehta. 2010. "Learning the Importance of Latent Topics to Discover Highly Influential News Items." In KI 2010: Advances in Artificial Intelligence, edited by Rüdiger Dillmann, Jürgen Beyrer, Uwe D. Hanebeck and Tanja Schultz, 211–218. Berlin: Springer.

Luotolahti, Juhani, Kanerva, Jenna, Laippala, Veronika, Pyysalo, Sampo, Ginter, Filip 2015. Towards Universal Web Parsebanks. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015).*

Laihiala, Tuomo and Ohisalo, Maria. 2017. Sosiaalipummit leipäjonossa? Kansalaisten käsityksiä huono-osaisten ansaitsevuudesta. [Welfare Parasites in Food Aid Line. Deservingness perceptions of general public]. In: Saari, J. (Ed.), *Sosiaaliturvariippuvuus: Sosiaalipummit oleskeluyhteiskunnassa?* Tampere: Tampere University Press, 233–258.

Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S. and Dean, Jeff. 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111–3119.

Niemelä, Mikko. 2008. Perceptions of the Causes of Poverty in Finland. *Acta Sociologica* 51 (1), 23-40.

Rehurek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop: New Challenges for NLP Frameworks. Valletta: University of Malta. 46–50.

Roberts, Margaret E. and Stewart, Brandon M. and Airoldi, Edoardo M. 2016a. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association,* 111 (515), 988–1003.

Roberts, Margaret E. and Stewart, Brandon M. and Tingley, Dustin. 2016b. Navigating the local modes of big data: The case of topic models. In: Alvarez, Michael R. (Ed.), *Computational Social Science: Discovery and Prediction.* New York: Cambridge University Press, 51–97.

Roberts, Margaret E. and Stewart, Brandon M. and Tingley, Dustin and Airoldi, Edoardo M. 2013. The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Saari, Juho, Behm, Miia and Lagus, Krista. 2017. Sosiaalipummi! – Moraalipaniikki 2010-luvun Suomessa. [Welfare bum! - Moral panic in Finland in 2010. In: Saari, J. (Ed.), *Sosiaaliturvariippuvuus: Sosiaalipummit oleskeluyhteiskunnassa?* Tampere: Tampere University Press, 207-231.

Suur-Askola, Laura-Maija. 2017. *”Perusoletus lienee se, että köyhä on laiska, tyhmä ja epärehellinen.” – Korpusavusteinen diskurssintutkimus köyhä- sanan semanttisesta prosodiasta Suomi24-keskustelupalstalla.* [“The basic assumption seems to be that a poor is lazy, stupid and dishonest”. A corpus-assisted discourse study on the semantic prosody of *poor* in the Suomi24 discussion forum.] BA thesis, University of Jyväskylä. https://jyx.jyu.fi/dspace/handle/123456789/53784.

Van Oorschot, W. 2006. Making the difference in social Europe: deservingness perceptions among citizens of European welfare states. *Journal of European Social Policy* 16 (1), 23–42.