- Contentiousness is about perspective, as seen in the complex patterns of agreement above. Based on these patterns, and identity variables about the participants, can we recognise and analyse different perspectives?
- To what extent can ConConCor be used to investigate bias in other types of data such as other cultural institutions' collections or contemporary newspapers?
- Can contentiousness be modelled in structured resources such as Knowledge Graphs to reason over it?

6 CONCLUSION

In this paper, we presented ConConCor: the Contentious terms in Context Corpus. A set of 2,715 unique samples from historical Dutch newspapers annotated with information on whether a particular target term in the context is contentious or not according to at least 7 annotators for most of our samples. The corpus contains two sets of annotations: 150 unique samples annotated by domain experts, and 2,570 unique samples annotated by crowd workers for a total of 21,800 annotations. We analysed the corpus along the dimensions of inter-annotator agreement, term statistics and context statistics.

Regarding research question 1, we see that while the overall inter-annotator agreement appears low, a major part of the samples in the corpus is annotated with a high percentage agreement between annotators. This stresses the value of having a large number of annotators per sample. Agreement among crowd workers is lower than among experts, but can be improved by filtering out underperforming annotators. An analysis of the annotations shows that in many cases, contentiousness is not a property of a word itself, but depends on the context it appears in: most words are in some contexts deemed contentious and in others non-contentious. There are, however, also words that are (non-)contentious regardless of context, or that seem to be a matter of personal opinion.

Research questions 2 and 3 concern the automatic detection of contentiousness: Our analyses show that contentiousness is a complex concept, making automatic detection difficult outside of context. We do, however, find clear signals in terms' semantics that relate to contentiousness. In answer to research question 3, we find that hierarchical clustering can identify certain concepts that seem indicative of contentiousness or non-contentiousness, which could be further developed in a tool to explore the contentiousness dimensions of galleries, libraries, archives, and museum collections.

With this research line, we hope to make cultural heritage collections more sensitive to different societal tendencies, leading to more diverse and inclusive collection descriptions.

AUTHOR CONTRIBUTIONS

RB: Stratified sampling of snippets from the Europeana newspaper Archive, building the annotation UI, coordinating the Prolific task, and compilation of the dataset. Main author of the ConConCor datasheet. Co-author of section 4.2. AN: Inter-annotator agreement analysis and data visualisation (Section 4.1), corpus and code documentation, contributed to the Introduction, Related work, and Creating ConConCor sections. VV: Implementation of text snippet sampling and of computational analyses, co-author of section 4.2. JvO: Introduction, overall manuscript editing. LH: Conceptual design, introduction, analysis of inter-annotator agreement, conclusions. MvE: Conceptual design, project coordination, related work, outlook, conclusions & overall manuscript editing.

ACKNOWLEDGMENTS

This work was funded by the EuropeanaTech Challenge for Europeana Artificial Intelligence and Machine Learning datasets, 'Culturally Aware AI' funded by NWO, and SABIO funded by the Dutch Digital Heritage Network. The authors would like to thank the Cultural AI Lab and KNAW HuC colleagues for their comments and annotations and the anonymous Prolific annotators. Special thanks to Mirjam Cuper (National Library of the Netherlands) for guiding KB and Europeana procedures, Lynda Hardman (CWI) for the suggestions on the article editing, and the anonymous reviewers for their constructive feedback.

REFERENCES

[1] Csilla E. Ariese. 2020. Amplifying Voices: Engaging and Disengaging with Colonial Pasts in Amsterdam. *Heritage & Society* 13, 1-2 (March 2020), 117–142. https://doi.org/10.1080/2159032X.2021.1901335

[2] Akanksha Bisht, Annapurna Singh, H. S. Bhadauria, Jitendra Virmani, and Kriti. 2020. Detection of hate speech and offensive language in twitter data using LSTM model. *Advances in Intelligent Systems and Computing* 1124 (March 2020), 243–264. https://doi.org/10.1007/978-981-15-2740-1_17

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[4] Rosagemma Ciliberti, E. Fulcheri, Paolo Petralia, and Anna Siri. 2021. Sharing ethics of displaying human remains in museums. *Medicina Historica* 4, 3 (Feb. 2021), 1–8. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101736438&partnerID=40&md5=c0d03a2b83637982387afaf4437062ae

[5] Carol Ann Dixon. 2012. Decolonising the museum: Cité Nationale de l'Histoire de l'Immigration. *Race & Class* 53, 4 (March 2012), 78–86. https://doi.org/10.1177/0306396811433115

[6] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. (2018). arXiv:1808.06080

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for Datasets. (2020). arXiv:1803.09010

[8] Kateryna Kaplun, Christopher Leberknight, and Anna Feldman. 2018. Controversy and Sentiment: An Exploratory Study. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (Patras, Greece) *(SETN '18)*. Association for Computing Machinery, 1–7. https://doi.org/10.1145/3200947.3201016

[9] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

[10] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (Sept. 2018), 861. https://doi.org/10.21105/joss.00861

[11] Wayne Modest and Robin Lelijveld. 2018. *Words Matter: an unfinished guide to word choices in the cultural sector*. Technical Report. The National Museum for World Cultures (Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum). https://www.tropenmuseum.nl/en/about-tropenmuseum/words-matter-publication

[12] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (Valletta, Malta). University of Malta, 46–50. https://is.muni.cz/publication/884893/en

[13] Marcus Schulzke. 2012. Contentious Language: South Park and the Transformation of Meaning. *Journal of Popular Film and Television* 40, 1 (March 2012), 22–31. https://doi.org/10.1080/01956051.2011.624136

[14] Brenda Trofanenko and Avner Segall. 2014. *Beyond Pedagogy: Reconsidering the Public Purpose of Museums. Introduction*. SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6209-632-5_1

[15] Johanna Turunen and Mari Viita-Aho. 2020. Changing interpretations of the Gallen-Kallela Museum's Africa collection. *Historiallinen Aikakauskirja* 118, 4 (2020), 466–480. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098751112&partnerID=40&md5=796942175e7a5a6bcf711bef454e13ba