DIVERSITY AND BIAS IN DBPEDIA AND WIKIDATA AS CHALLENGES FOR DOWNSTREAM PROCESSING

Bettina Berendt
TU Berlin &
Weizenbaum Institute,
Germany,
KU Leuven, Belgium

Özgür Karadeniz
KU Leuven,
Belgium

Sercan Kıyak
KU Leuven,
Belgium

Stefan Mertens
KU Leuven,
Belgium

Leen d'Haenens
KU Leuven,
Belgium

Abstract

In this research, we compare three data sources that our text analysis tool Diversity Searcher has worked with – DBpedia in two languages and Wikidata – with respect to their ontological coverage and diversity, and describe implications for the resulting analyses of text corpora. We describe a case study of the representation of Belgian political parties between 1990 and 2020. In particular, we found a staggering overrepresentation of the political right in the English-language DBpedia.

Keywords: diversity, bias, automated text analysis, DBpedia, Wikidata

Introduction

Diversity Searcher (DS) is a semi-automated text analysis and knowledge enrichment tool designed to present information to the user about the degree of diversity in news media texts. The tool was developed in the interdisciplinary project DIAMOND (Diversity and Information Media: New Tools for a Multifaceted Public Debate)[1] to be used by media professionals and media consumers, and is to be integrated with iCandid,[2] an online platform offering integrated access to several (social) media resources to researchers with a focus on the social sciences and the humanities. DS relies on external knowledge sources to identify actors in media texts and to retrieve relevant properties and relationships between them, which brings the disadvantage of including these sources' errors and biases.[3] This led us to question the implications of our initial choice of the English-language DBpedia and compare it with the Dutch-language DBpedia and Wikidata.[4]

Methods

We concentrated on the attributes "political party affiliation"[5] and "political alignment of a party"[6] and focused our analysis on "politicians from Belgium" to ensure sufficient domain knowledge for steering and interpreting the analysis. As a baseline, we used the shares of the vote or the number of seats of parties at times T in the national parliament, and also looked at the Flemish parliaments.[7] We studied this in five-year intervals starting in 1990.[8]

Interpreting the ontologies as cultural memory, we asked what image the ontologies (in their current form) give of the representation of parties at these time points in the past. We queried the ontologies with SPARQL and postprocessed the data to obtain lists of all represented politicians active at the studied time points T. We regard these as giving visibility to the party or parties they belonged to. Individuals with multiple party affiliations across their career may be perceived differently (giving visibility to all of these parties, or to only one of them, or re-centring attention on themselves and thus not giving visibility to any party). We therefore

[1] KU Leuven Institute for Media Studies: Diamond, https://soc.kuleuven.be/fsw/diamond/, last accessed 11.03.2023.

[2] KU Leuven Libraries: A snapshot of LIBIS research infrastructures, https://bib.kuleuven.be/english/libis/projects#icandid, last accessed 11.03.2023.

[3] DBpedias and Wikidata in turn depend on other sources, processes (Wikipedias, extraction algorithms, and diverse datasets) and their biases; we do not investigate such upstream effects here.

[4] More details about Diversity Searcher as the context of this work can be found in Berendt et al., 2023.

[5] obtained via the attributes with highest coverage: dbp:party (EN), dbpedia-owl:party (NL), wdt:P102 (WD)

[6] To obtain complete coverage, we used dbp:align (EN), dbpedia-owl:align (NL), wdt:P1387 (WD), and postprocessed manually.

[7] https://nl.wikipedia.org/w/index.php?title=Kamer_van_volksvertegenwoordigers&oldid=61849646, https://nl.wikipedia.org/w/index.php?title=Vlaams_Parlement&oldid=61662474

[8] T = 1 January of 1990, 1996, 2000, 2005, 2011, 2015, 2020. The exceptions from the 5-year spacing were done to capture the effects of the general elections held in 1995 and 2010.

derived a lower bound on the visibility of any given party (politicians who only ever belonged to that party) and an upper bound (politicians who belonged to that party and possibly also to others).

Results

Figures 1 and 2 show examples of over- and under-representation: the 2015 data for the English-language DBpedia (Fig. 1 a), the Dutch-language DBpedia (Fig. 2 a) and Wikidata (Fig. 1 b). In each diagram, the parties are ordered from left to right according to their ideology (and then according to their acronym). For example, figure 1a shows that the 52 politicians associated with the right-wing N-VA constituted more than 60% of the 2015 politicians in the English DBpedia (bold line), while the party had obtained only 22% in the 2014 elections (thin line). The results confirm our informal observation of over-representation of right-wing parties (especially the N-VA) in the English-language DBpedia. Different biases seem to occur in the Dutch-language DBpedia: although these data are overall similar to the baseline, the main centrist party seems overrepresented (CD&V with close to 20% of 355 represented politicians compared to 12% of the national vote). Wikidata, in contrast, gives a rather accurate picture of party shares in the national parliament. Similar overrepresentations in media coverage have been identified in earlier international research, such as centrist bias in media coverage of the UK elections of 2017 (Deacon et al., 2017) and right-wing overrepresentation in social media, despite cries of censorship in the United States (Mark, 2020).

The seeming overrepresentation of the political centre in the Dutch-language DBpedia may be an artefact of how language, political system, and data creation interact. Figure 2 (a) also shows that the French-language Walloon parties (esp. Ecolo, PS, Les Engagés) are under-represented. Figure 2 (b) maps the shares of seats in the Flemish parliament (which was first elected directly in 1995) as the baseline and is otherwise analogous to the other figures. The set of parties is a subset, since only the Flemish parties can be elected to the Flemish parliament (while the national parliament also contains representatives from the Walloon, Brussels, and German-Community parties). The contrast between Fig. 2 (a) and (b) suggests that the representation of the parties between extreme left and centre-right in the Dutch-language DBpedia mirrors the shares of these parties in the regional parliament rather closely; while the right and extreme right tend to be underrepresented especially in the later years, when they were highly successful, especially in the Flemish elections.

Figures (including in larger size) and interactive figures for all years 1990 to 2020 are available, which also illustrate that the overrepresentation of the political right in the English-language DBpedia increased over time.[9]

Discussion and Conclusions

Data analyses such as this one alone cannot answer the “why” of biased representations, and this “why” requires further research. Regardless of reasons, however, our results highlight the need to remain aware that data not neutral and working with them requires applying knowledge and situating them in a historical and social context.

The aforementioned/above problem can be exacerbated by the inability of automated tools to recognise and understand context. This is another reason to treat the numerical and categorical results as a starting point for deeper text analysis and involve users in sense-making – whether they are an individual citizen, researcher, organisation representing a particular target group or a journalist looking for a new angle for a news story. For future work, a two-pronged strategy is recommended: (a) identifying and using the best-suited ontology and data for a given task and (b) making its properties and shortcomings transparent to users so as to keep users aware of challenges associated with the (and any) dataset.

Acknowledgements

We thank the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO) for funding DIAMOND under project code S008817N.

References

Berendt, B.; Karadeniz, Ö.; Kıyak, S.; Mertens, S.; d'Haenens, L. Diversity and bias in DBpedia and Wikidata as a challenge for text-analysis tools. O-bib. Das offene Bibliotheksjournal, 10(2). 2023. https://doi.org/10.5282/o-bib/589

Deacon, D.; Downey, J.; Smith, D., Stanyer, J.; Wring, D. National News Media Coverage of the 2017 election. Centre for Research in Communication and Culture, Loughborough University Report 4: 5 May – 7 June 2017, https://blog.lboro.ac.uk/crcc/wp-content/uploads/sites/23/2017/06/media-coverage-of-the-2017-general-election-campaign-report-4.pdf

Scott, M.: Despite cries of censorship, conservatives dominate social media, POLITICO, 26.10.2020, https://www.politico.com/news/2020/10/26/censorship-conservatives-social-media-432643

Karadeniz, Ö.; Berendt, B.; Kıyak, S.; Mertens, S.; d'Haenens, L. Political representation bias in DBpedia and Wikidata as a challenge for downstream processing, arxiv.org, 2022. https://doi.org/10.48550/arXiv.2301.00671

[9] They can be found in Karadeniz et al., 2022 and at http://www.berendt.de/DIAMOND/.

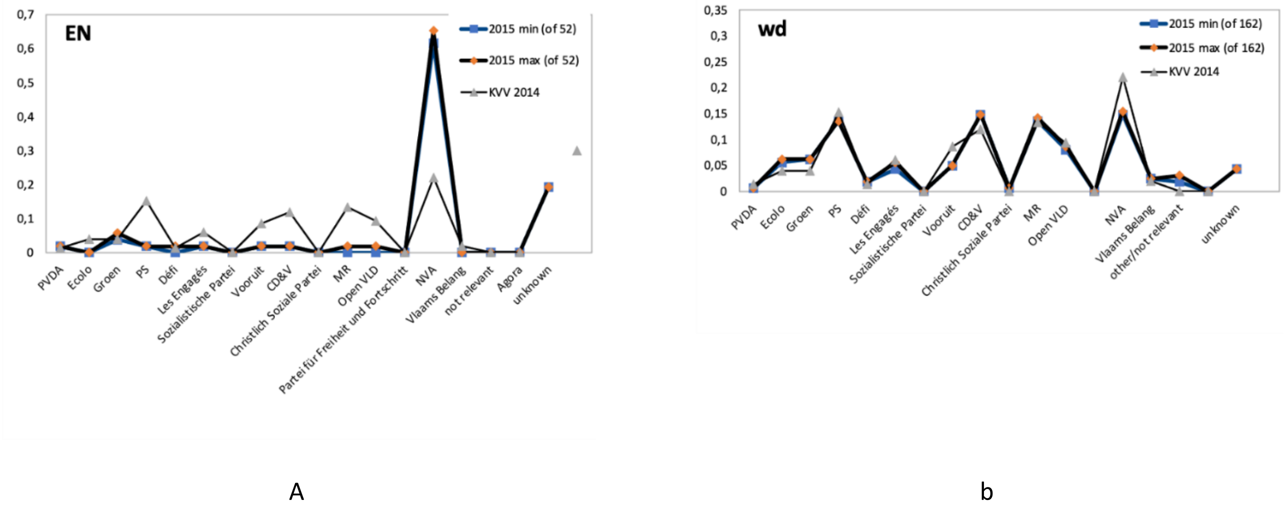

Figure 1: Visibility of parties in the English-language DBpedia (a) and Wikidata (b) for 2015. Parties are ordered along the political alignment (left to right). The bold lines are the proportions of representation on the database (upper and lower bounds); the thin line (KVV) shows the real proportion of the votes obtained. Proportions range from 0 to 1 (~ 100%).

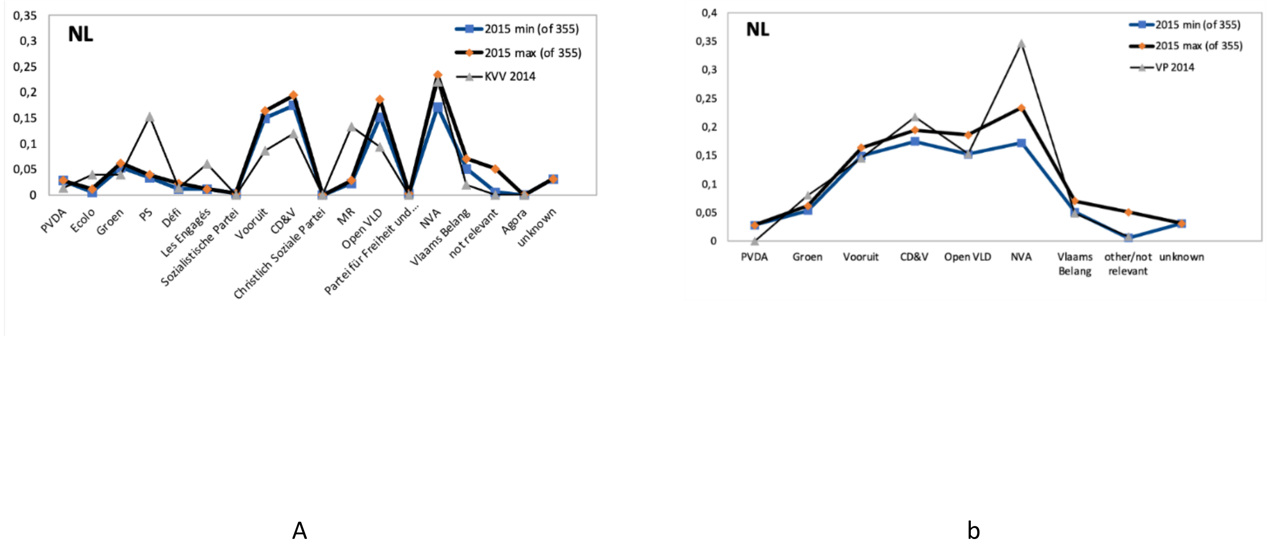

Figure 2: Visibility of (a) all parties and (b) the Flemish parties in the Dutch-language DBpedia for 2015. The bold lines are the proportion of representation on the database (upper and lower bounds); the thin grey lines (KVV resp. VP) show the real proportion of the votes obtained. X and Y axis are ordered as in Figure 1.