Linking Survey and Social Media Data: Natural Language Processing for Bridging the Gap Between Open Access and Data Protection

Conor Gaughan, Rachel Gibson & Alexandru Cernat *
Cathie Marsh Institute for Social Research (CMI)
University of Manchester
Manchester, UK
{conor.gaughan,rachel.gibson,alexandru.cernat}@manchester.ac.uk

Marta Cantijoch Cunill
School of Social Sciences
University of Manchester
Manchester, UK
{marta.cantijoch}@manchester.ac.uk

Riza Batista-Navarro
Department of Computer Science
University of Manchester
Manchester, UK
{riza.batista}@manchester.ac.uk

Abstract

The open release of social media data is problematic for both ethical and legal reasons, and the publicly searchable nature of social media text imposes a serious risk of disclosure. This is especially risky when linking social media data with participant survey responses which will likely contain sensitive information. This work-in-progress paper seeks to outline a standardised procedure for the extraction of anonymised variables from participant social media data which can be safely shared with other researchers. Using two pre-existing datasets which link participant survey data with their X (formerly Twitter) profiles during the US 2020 and 2024 elections campaigns, we use NLP methods to extract 126 variables which describe the structural and semantic nature of the social media text. Doing so, we look to demonstrate how these new variables can be used to enhance public opinion research such as the prediction of socio-demographic and attitudinal characteristics.

1 Introduction

Public opinion research is entering a new era of development that centres on the linkage of conventional survey data with other forms of observed or digitally generated data (Stier et al., 2020). This so-called "third era" of public opinion research has been marked by the growing integration of new technologies such as web trackers, mobile apps, and digital sensors into survey research (Groves, 2011). Commonly referred to as digital trace data (DTD), this can include communication data, web browsing data, geo-location data, and data extracted from the Internet and various social media platforms. These new digital technologies have unlocked entirely new areas of study for researchers which conventional survey data cannot reach. However, much DTD – especially publicly searchable social

*This work has been conducted as part of the UKRI Smart Data Research UK funded project DIGISURVOR. Project webpage can be found here: https://digisurvor.github.io/main/

media data (SMD) – is identifiable and makes the open sharing of such data difficult to do. However, open sharing of research data is fundamental to the principles of FAIR data: that is, increasing the public value of these datasets by making them more findable, accessible, interoperable and reuseable (Wilkinson et al., 2016). Thus, this working paper seeks to bridge the gap between open access and data protection by proposing a standardised and replicable process for the representation of information drawn from SMD, that can prevent disclosure while maintaining utility. The said process harnesses natural language processing (NLP) methods for extracting variables pertaining to public opinion, which include sentiments and emotions.

2 Data and Methods

We employ a range of computational and NLP approaches to derive new variables of interest from our SMD which can be safely augmented to participant survey data and securely shared with the wider research community. We employ state-of-the-art, large language models (LLMs) to extract variables from participant SMD text fields to broadly describe and characterise their general online behaviour. When linked with their survey responses, this can unlock new insights into human behaviour and public opinion. To demonstrate that SMD can in fact be effectively made available for open access while maintaining both privacy and utility, we make use of two preexisting datasets from X (formerly Twitter). We build upon linked survey-to-X datasets for two Presidential election time points in the United States (US) – 2020 (1) and 2024 (2).[1] The survey component measures media consumption, perceptions of digital campaign contact, awareness of misinformation, core political attitudes and behaviours, plus standard socio-demographic characteristics and self-reported X use. Participants were then invited to share their X handle with the research team, which was subsequently used to extract the timelines, follows and likes of these respondents, as well as the collected the timelines of the accounts that respondents were following.

Dataset 1 was fielded to 5,952 respondents between 16 September – 20 October 2020, with a second follow-up wave fielded in the week following the election (9 November). 2,460 participants reported having an active X account (41%) and 1,598 agreed to share their handle (27%). 920 of these were successfully validated against the X API, which constituted 15% of the overall sample. Data extraction from X was restricted to the election period 1 September 2020 – 9 November 2020. Dataset 2 was fielded to 5,757 respondents between X and Y. 2,621 reported having an active X account (46%) and 1,570 agreed to share their handle (27%). 963 of these were successfully validated against the X API which constituted 17% of the overall sample.

Our selection of derived variables from the linked X data is initially guided by the previously published datasets of research bodies also working with linked survey-to-social media data: namely, Understanding Society UK, GESIS, and the American National Election Studies (ANES). These have involved the extraction of various lexical, syntactic and grammatical features from social media text such as standard (e.g. letters, sentences, syllables) and special (e.g. emojis, hashtags, @mentions) character counts, to more advanced measures of grammaticality, readability, and sentimentality. Aligned with previous work showing NLP can support public opinion research (Callegaro & Yang, 2017; Karamouzas et al., 2022)), we aim to build on these existing procedures to demonstrate the way in which large language models (LLMs) can be harnessed to improve the openness and availability of SMD used in research.

Selecting a range of structural and substantive variables derived from participant X profile fields (display names, descriptions, locations, public metrics) and timeline posts (text, engagement metrics), we draw on a suite of NLP models from various sources. We use Python's textstat and LanguageTool libraries for computation of language readability, grammaticality, and sophistication, Google's Perspective API for toxicity classification (incl. detection of threats, identity attacks, and profanity), and TweetNLP (Camacho-Collados

[1] Both were originally collected by Prof. Rachel Gibson for the European Research Council funded Project 833177(2020- 2025) Digital Campaigning and Electoral Democracy (DiCED).

et al., 2022) for topic classification, irony detection, hate speech detection, sentiment analysis and named entity recognition. Additional models such as deepIdeology (Gottlieb, 2018) can also help us to measure ideological position of individual posts. These are combined with geo-location matching, real name certification and URL domain matching. These variables have all be carefully selected to provide researchers with quantifiable measures of what participants are posting on their profiles and how they are doing it, while avoiding disclosure of any directly identifiable information.

3 Early Findings and Ongoing Work

We have thus far generated a total of 126 variables split between profile-level (65) and post-level (61) metadata. Our next steps are to prove that these derived variables can adequately meet two essential targets: (1) effective anonymity; and (2) effective utility.

3.1 Demonstration of Utility

As proof of utility, we will demonstrate the maintained value of these new variables by exploring several substantive research questions using our newly linked survey-to-SMD. For example, one of the ways in which we can leverage NLP and linked social media datasets to improve public opinion research is the potential to predict survey responses using the augmented variables. This could help to improve survey research by using the newly derived variables to impute missing values in key survey variables. To illustrate, we select a subset of key socio-demographics and attitudinal variables from the respondent survey data [age, gender, education level, household income, ideological position, political attention] as well as the following twelve social media variables: % of tweets that contain media, mean tweet engagement, mean post length, the mean number of capital letters, hashtags, emojis, and URLs used in their tweets, their mean grammar and readability scores, the % of tweets with neutral sentiment, and the % of offensive tweets. We also include the % of tweets that talk about each of the top five topics: (1) news and social concern, (2) dairies and daily life, (3) sports, (4) film, TV and radio, and (5) business and entrepreneurship.

The total number of individual tweets in the US 2020 dataset is 222,365. After dropping participants who had posted less than ten tweets and dropping rows with missing data, we are left with 259 participants in the US 2020 data. Splitting the data into train (80%) and test (20%) sets, we fit seven binary/ordinal logistic regressions. Using these models to predict the out-of-sample test data, our subset of derived variables was able to successfully predict respondent age, gender and household income significantly better than with no information at all (See Figure 1):

As proof of anonymity, we will conduct various anonymity tests, following the statistical disclosure control (SDC) framework. This incudes k-anonymity, t-closeness and l-diversity testing as well as a global risk assessment using the *sdcMicro* library in R.

4 Limitations

When replacing raw text with NLP-derived aggregated features, this creates a shortcoming in terms of information loss. There is a direct trade-off between the need to protect against disclosure by not releasing the raw data and the aim to encourage open research by releasing anonymous extracted variables from the data in lieu. We acknowledge this limitation and accept that to openly provide some utility to researchers from our linked datasets without compromising confidentiality, a degree of information loss is necessary. We have deliberately selected open source and clearly documented NLP models for the benefits of transparency and replicability, so that other researchers using these variables will know exactly how they have been derived and can assess their quality. Also, while our original survey datasets contained over 5,000 participants each, various stages of non-response bias (1. no X account; 2. no consent to linkage; 3. failure to link; 4. no tweets) means that our final linked samples are relatively small. Additionally, due to selection effects ate every stage of non-response, the final samples are also unlikely to be nationally representative. Nonetheless, we believe

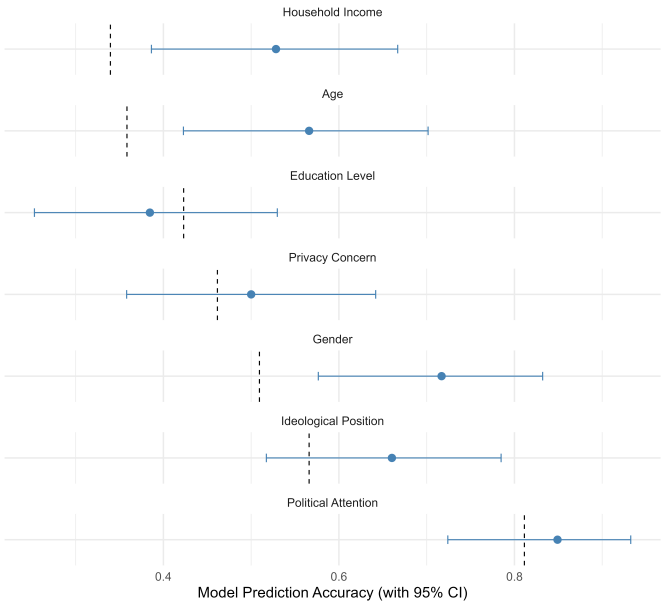

Figure 1: Confusion matrix coefficients showing the accuracy of each model with 95% confidence intervals compared to the No Information Rate (NIR) baseline (dashed black line). Models for age, gender and household income performed significantly better than with no information at all

this is a good starting point for recognising the potential of using NLP methods for the benefit of public opinion research.

5 Conclusion and Next Steps

This work-in-progress paper looks to establish a standardised procedure for extracting variables from social media data which can be made open access. We seek to harness the power of modern NLP techniques and LLMs to improve the sharing of linked survey-to-SMD within the scientific community and enhance the reusability of SMD and other forms of DTD. We do this by employing various types of NLP models to extract variables of interest such as sentiments, topics, grammaticality, readability and entity recognition. While the underlying textual data cannot be shared in its raw form due to the publicly searchable nature of SMD, this work will demonstrate the effectiveness of NLP in deriving structural and substantive variables from these data which can still be useful. As several scholars have pointed out, the eagerness of researchers to mass collect data from social media platforms has vastly outpaced the adaptation of our traditional ethical frameworks in public opinion research (Williams et al., 2017; Sloan et al., 2020). Our aim is to begin the process for addressing such concerns and to highlight the positive impact that modern developments in NLP can have for improving standards in public opinion research.

References

Mario Callegaro and Yongwei Yang. The role of surveys in the era of "big data". In *The Palgrave handbook of survey research*, pp. 175–192. Springer, 2017. doi: https://doi.org/10.1007/978-3-319-54395-6_23.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*, 2022. doi: https://doi.org/10.48550/arXiv.2206.14774.

Alex Gottlieb. deepideology: Scale ideological slant of tweets. In *GitHub Repository*, 2018. URL https://github.com/alex-gottlieb/deepIdeology.

Robert M Groves. Three eras of survey research. *Public opinion quarterly*, 75(5):861–871, 2011. doi: https://doi.org/10.1093/poq/nfr057.

Dionysios Karamouzas, Ioannis Mademlis, and Ioannis Pitas. Public opinion monitoring through collective semantic analysis of tweets. *Social Network Analysis and Mining*, 12(1): 91, 2022. doi: https://doi.org/10.1007/s13278-022-00922-8.

Luke Sloan, Curtis Jessop, Tarek Al Baghal, and Matthew Williams. Linking survey and twitter data: informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, 15(1-2):63–76, 2020. doi: https://doi.org/10.1177/1556264619853447.

Sebastian Stier, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5):503–516, 2020. doi: https://doi.org/10.1177/0894439319843669.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. doi: https://doi.org/10.1038/sdata.2016.18.

Matthew L Williams, Pete Burnap, and Luke Sloan. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168, 2017. doi: https://doi.org/10.1177/0038038517708140.