MI

Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences

Nicole Schwitter[1,2] (iD)

Abstract
Recent months have witnessed an increase in suggested applications for large language models (LLMs) in the social sciences. This proof-of-concept paper explores the use of LLMs to improve text quality and to extract predefined information from unstructured text. The study showcases promising results with an example focussed on historical newspapers and highlights the effectiveness of LLMs in correcting errors in the parsed text and in accurately extracting specified information. By leveraging the capabilities of LLMs in these straightforward, instruction-based tasks, this research note demonstrates their potential to improve on the efficiency and accuracy of text analysis workflows. The ongoing development of LLMs and the emergence of robust open-source options underscores their increasing accessibility for both, the quantitative and qualitative, social sciences and other disciplines working with text data.

Keywords
Text analysis, optical character recognition, large language models, information extraction, data quality

Introduction

Recent months have witnessed a surge in suggested applications for large language models (LLMs) in the social sciences: LLMs are being used to generate synthetic survey answers (Argyle et al., 2023), as chatbots in conversations in intervention studies (Costello et al., 2024), or to detect constructs and annotate and label text in a variety of contexts (Macanovic and Przepiorka, 2024; Rathje et al., 2023; Ziems et al., 2024). As one of the first, Brown et al. (2020) have argued that LLMs are few (or zero) shot-learners and can greatly improve performance in many of the typical natural language processing (NLP) tasks.

These uses of LLMs come with their limitations and critique (see also Brown et al., 2020). The deep neural networks used in LLMs are widely recognised as black boxes with opaque decision-making processes (Dobson, 2023) which retain the biases inherent in the data that they had

been trained on (Bender et al., 2021; Navigli et al., 2023). Also, asking LLMs the right question is not that straightforward. 'Prompt engineering', the crafting of efficient questions, is an iterative process of trial and error, guided by best practices and further complicated by the fact that LLMs are sensitive to changes in wording and not necessarily deterministic so that the same prompts can lead to different responses (Chen et al., 2023; Ouyang et al., 2023; Strobelt et al., 2023). Given these difficulties regarding reproducibility, the usefulness of LLMs for (social)

[1]Mannheimer Zentrum für Europäische Sozialforschung, University of Mannheim, Mannheim, Germany
[2]Department of Sociology, University of Warwick, Coventry, UK

Corresponding author:
Nicole Schwitter, Mannheimer Zentrum für Europäische Sozialforschung, University of Mannheim, A5, 6 (Department A), Mannheim 68159, Germany.
Email: Nicole.schwitter@uni-mannheim.de

scientific use cases is an important topic of debate (Ball, 2023; Ye et al., 2023).

This research note will highlight one of LLMs' strengths amidst their acknowledged shortcomings: Their ability to comprehend even poorly formatted or low-quality text and to extract predefined information from it. Text-as-data has amassed within the last decade. Text-as-data first undergoes a pipeline of different preprocessing steps, which include steps like tokenisation, the removal of stopwords or stemming (e.g. Grimmer et al., 2022: 50–53). Effective text analysis hinges on the quality and cleanliness of the input data, but the reality often looks messy: Social media data, for example, is user-generated and can thus exhibit all kinds of errors and misspellings. Digitised, optical character recognised (OCR-ed) text data is also often plagued by inaccuracies and artifacts introduced during the digitisation process. Unlike other tools, LLMs circumvent the need for extensive data cleaning and preprocessing and can instead be advised to take over this task. In addition, LLMs are well equipped to extract pieces of information from text. Through a proof-of-concept demonstration, this research note will discuss how LLMs can effectively be integrated into the research process when working with unstructured text data.

Improving the text analysis workflow with large language models

From open-ended survey questions to comments scraped from social media pages and digitised books, text data has amassed within the last decade. These data sources hold information in unstructured text instead of tidy numeric datasets, making them less straightforward to work with. This increase in data has led to the development of more sophisticated methods to systematically and automatically understand and process texts as data (Grimmer and Stewart, 2013).

Already before the increased popularity of LLMs in the recent years, language models have been suggested to improve untidy text (e. g. Neto et al., 2020; Xu et al., 2022). Given the new widespread and transdisciplinary interest in LLMs, this research note advocates for their inclusion in the toolbox of social scientists who work with text data. LLMs are advanced artificial intelligence systems designed to interpret and generate human language. They are based on neural networks and are trained on vast amounts of textual data which has allowed them to learn patterns, semantics and grammar.

One important emergent capacity of these models is that they can analyse textual statements in line with a question posed by the user, including typical use cases in the social sciences like identifying themes and emotions, or coding the text in terms of key features like hate speech or other labels, with new tasks still being discovered

(Lupo et al., 2023; Ornstein et al., 2023; Törnberg, 2023a, 2023b). However, as users pose more complex questions, granting the model greater autonomy and agency in its responses (see also Latour, 1996), the risk of introducing biases and limiting replicability also increases. With more complex questions, LLMs base their answers on more contextual knowledge; however, the contextual knowledge at play is unknown to the researcher. LLMs do not offer a positionality statement (Bourke, 2014) – as researchers and users, we do not have a clear understanding of underlying biases, assumptions and values guiding the model's responses, so that it becomes challenging to discern the basis upon which decisions are made.

On the other hand, LLMs excel in tasks where they are provided with narrow instructions and have limited autonomy; these tasks are often more straightforward but can still be challenging to automate in research contexts. These tasks include, for example, correcting messy OCR text or misspelled social media posts, or extracting specific pieces of information. In these scenarios, the model's capacity to understand and follow precise instructions allows it to perform with accuracy and efficiency, while still being more flexible than more traditional approaches (such as named entity recognition) as they manage to understand unclear instructions and input text. As Törnberg (2023b) phrased it, LLMs can be thought of as virtual student assistants instructed with textual analysis which are versatile and capable, but prone to misunderstandings - narrow instructions and clear use cases minimise these misunderstandings. For instance, when tasked with correcting OCR or spelling mistakes, LLMs can leverage their language understanding capabilities to identify and rectify errors, improving the overall quality and readability of the text. Similarly, when instructed to extract specific information, LLMs can navigate through the text, locate relevant details and extract the requested information.

An example: Preprocessing historical newspapers with Command R +

To showcase the capabilities of an LLM, I take historical newspapers excerpts as an example. In the following, I aim to retrieve the bride's and groom's names from American newspapers from 1861. The digitised newspapers from *Chronicling America* are available as OCR-ed text; the data quality of these texts varies. Information extraction with regular expressions and named entities becomes particularly challenging as there are diverse phrasings of wedding reports (e.g. 'Miss X gets married to Mr X', 'the unity of Mr X and Miss X', 'Miss X, daughter of Y, celebrated her wedding to Mr X, son of Z', etc.), they often include misidentified letters (e.g. 'Miss' can easily be parsed as 'Mlss') and line breaks in the original

newspaper text has introduced hyphenations which are difficult to resolve.

Example text look as follows:

Excerpt 1:'Married?On Wednesday, the 18th Instant, by Rev. J. M. H. Adam*, Capt. R. LEANDER TOMLINSON* and Mist MARTHA S. WRIGHT, all of this District. We, of the Enquirer office, return our prettiest thank* to the pnrtles for a large portion of the wedding cake. May their path of life be unobstructed by cares and griefs.' (National Endowment for the Humanities, 1861d)

Excerpt 2:'Married. Oa the 19th ult., by the Rev. 0. H. Martin. Major Andrew Jackson Bosweix, to Miss Cynthia A. Jack–on, all of this couuty. The Major was married immediately ou the reception of his commission, which nives u.i tho satisfaution of kuow- w we niadn him a Major aud got bim a very handsome wifo. On the 20th ult., by tho Ilev. J. B McLelland, Mr. Geo. N. Ladd to Miaa Susan Murray, all of tbts oounty. 0 ! that all the ladies could get sucli lads as old man George. e placed a piece of bis wedding cake under our pillow, hoping to dream of a widowbut it made us dream of having ou our soi dier clothes. A suro sign ot war. Two ladies at the weddiug being m vited to favor the company with musio played, 'My hopo has departed forever which might lead one naturally to infe that old man Georgo was a gay deceiver' (National Endowment for the Humanities, 1861b)

Excerpt 3:'___ MARRIED—In Fulton county, Ark., July L'-tli, 1861, at the residence of the bride's lather, by Elder Butler, Mr. Alfred Turner to Miss 8aiiaii Vincent. Arkansas papers please copy. In Fulton couaty, Ark., Aug. 2d, IfeGl.at the residence of Hon. S. W. Cochran, by Rev. If my Livingston, Mr. Jas. McCarlky to Miss Lons v; Cochran, distant relative of Hon. S. W. Cothran.; Ark. and Georgia papers please copy.' (National Endowment for the Humanities, 1861a)

Excerpt 4:'In tbe Weekly paper ouly, or at irregular interval iu either of tbe, aaMfa, wttl b- charged al per enuare for each ami eiery in-er-lon. tW Anmiuncing Candida tea for Sia'e. County and Municipal office. $5 each, lo be pail in advance in every iimtmic. Marriage aud Death are pnblinhed a-i uews bot Obi-narie. TributeH of Itepect niid Funeral Iuvi talionn at. alt other advertiaevaenui. All legai and transient adeertlaement wl 1 be charged by the laaul Untj.' (National Endowment for the Humanities, 1861c)

These texts provide challenges for rigid approaches. However, the data can be preprocessed by prompting the LLM to improve the text quality and to extract specific information in a desired format. Via an application programming interface (API), the LLM can be accessed with a few lines of code by passing a prompt (see also Törnberg, 2023b). Different LLMs exist, the most popular one currently being GPT-4 (developed by OpenAI). I make use of open-source alternatives via HuggingFace and use the model Command R + , which is one of Cohere's LLMs and which was the first open-weight model to beat GPT4 in the Chatbot Arena, a crowdsourced open platform for LLM evaluations (as of the writing of this manuscript in April 2024).

Advising the LLM to correct the text (using the following prompt: 'You are an OCR expert. You are perfect at fixing errors which happen when digitising text. Please correct the following text and return the corrected version: [excerpt]') has led to the following results:

Excerpt 1:'Married—On Wednesday, the 18th instant, by Rev. J. M. H. Adams, Capt. R. Leander Tomlinson and Miss Martha S. Wright, all of this district. We, of the Enquirer office, return our prettiest thanks to the parties for a large portion of the wedding cake. May their path of life be unobstructed by cares and griefs.' (National Endowment for the Humanities, 1861d)

Excerpt 2:'Married. On the 19th ult., by the Rev. O. H. Martin, Major Andrew Jackson Boswell, to Miss Cynthia A. Jackson, all of this county. The Major was married immediately on the reception of his commission, which gives us the satisfaction of knowing we made him a Major and got him a very handsome wife. On the 20th ult., by the Rev. J. B. McLelland, Mr. Geo. N. Ladd to Miss Susan Murray, all of this county. Oh! that all the ladies could get such lads as old man George. We placed a piece of his wedding cake under our pillow, hoping to dream of a widow; but it made us dream of having on our soldier clothes. A sure sign of war. Two ladies at the wedding, being invited to favor the company with music, played, 'My hope has departed forever,' which might lead one naturally to infer that old man George was a gay deceiver.' (National Endowment for the Humanities, 1861b)

Excerpt 3:'Married—In Fulton county, Ark., July 15th, 1861, at the residence of the bride's father, by Elder Butler, Mr. Alfred Turner to Miss Sallie Vincent. Arkansas papers please copy. In Fulton county, Ark., Aug. 2d, 1861, at the residence of Hon. S. W. Cochran, by Rev. Jimmy Livingston, Mr. Jas. McCarley to Miss Loula Cochran, a distant relative of Hon. S. W. Cochran. Ark. and Georgia papers please copy.' (National Endowment for the Humanities, 1861a)

Excerpt 4:'In the Weekly paper only, or at irregular intervals in either of the dailies, will be charged at per square for each and every insertion. Announcing Candidates for State, County, and Municipal office, $5 each, to be paid in advance in every instance. Marriages and Deaths are published as news but Obituaries, Tributes of Respect, and Funeral Invitations are charged at regular advertising rates. All legal and transient advertisements will be charged by the square inch.' (National Endowment for the Humanities, 1861c)

Further prompting the LLM to return the name of the bride and groom as a comma-separated table (using the prompt: 'Please extract the name of the bride and groom and return it as a comma-separated table with two columns 'name_bride' and 'name_groom'. These are the people getting married and their first and last names are

given. If you do not find any names, put 'NA' in the table. Return nothing else.') led to the following output:

> **Excerpt 1:**'name_bride,name_groom
> Martha S. Wright,R. Leander Tomlinson.' (National Endowment for the Humanities, 1861d)
> **Excerpt 2:**'name_bride,name_groom
> Cynthia A. Jackson,Andrew Jackson Boswell
> Susan Murray,George N. Ladd' (National Endowment for the Humanities, 1861b)
> **Excerpt 3:**'name_bride,name_groom
> Sallie Vincent,Alfred Turner
> Loula Cochran,Jas. McCarley' (National Endowment for the Humanities, 1861a)
> **Excerpt 4:**'name_bride,name_groom
> NA,NA' (National Endowment for the Humanities, 1861c)

The LLM-processed text has cleaned errors introduced through OCR and has produced well-structured results. It is important to note that the results are not perfect; especially with difficult source material, improving the data is difficult (garbage in, garbage out). For example, the name '8aiiaii Vincent' in the third excerpt gets corrected to Sallie Vincent which is wrong; taking a look at the scanned image shows that the actual name is Sarah Vincent. Some errors thus still pertain, but the improvement in quality from the OCR-ed text is undeniable. Usage of the LLM has also allowed the extraction of key information which can now be analysed with more traditional methods of quantitative data analysis.

Conclusions

The (optimal) use of LLMs in research methodologies has emerged as a prominent subject in the recent scientific discourse. When passing LLMs clear instructions and well-defined tasks, the results become reproducible and simple to validate. In the presented example, I used Command R + to extract information from historic newspapers, leading to overall good results. Challenging source material led to less-than-ideal output results, but there is an impressive overall improvement in quality. Nevertheless, it is of utmost importance that results given by LLMs are validated. This proof-of-concept paper showed promising results for (English-language) newspapers, but the application of LLMs must be adapted and tested within specific research contexts and input texts. For example, considering untidy OCR, the task of correcting wrongly put letters (e.g. replacing l's with i's) is more feasible than correcting gibberish or wrong digits, and shorter inputs are generally easier to process than long texts (Chang et al., 2024).

Preprocessing and information extraction with LLMs will not always work and this also cannot be the benchmark, as the process of correcting errors is an equilibrium of correcting mistakes while not introducing many new

ones (Kim et al., 2021). While errors are thus part of the process, (open-source) LLMs have the capabilities to improve data quality and thus make more text accessible for research. This also means that text which is available in more niche contexts can be made more accessible, allowing for even more widespread use and analysis of text data within the social sciences. Improved open-source models like Command R + used in this study are democratising access to sophisticated language processing capabilities. As such, ongoing innovations and efforts to enhance the efficiency and affordability of LLMs hold promise for empowering research endeavours across diverse domains, including projects with limited resources (both financial and technical).

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Ethical approval was not required as only public data is being used.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

ORCID iD

Nicole Schwitter https://orcid.org/0000-0002-3837-680X

Data availability statement

Data used in this study is publicly available and cited.

References

Argyle LP, Busby EC, Fulda N, et al. (2023) Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3): 337–351.

Ball P (2023) Is AI leading to a reproducibility crisis in science? *Nature* 624(7990): 22–25.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: Can language models be too big?. In: *Proceedings of the 2021 ACM conference on fairness,*

accountability, and transparency, New York, NY, 1 March 2021, pp.610–623. New York, NY: ACM.

Bourke B (2014) Positionality: Reflecting on the research process. *The Qualitative Report* 19(33): 1–9.

Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. *arXiv*:2005.14165. (accessed 12 April 2024).

Chang Y, Lo K, Goyal T, et al. (2024) BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *arXiv*:2310.00785 (accessed 12 April 2024).

Chen B, Zhang Z, Langrené N, et al. (2023) Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review. *arXiv*:2310.14735 (accessed 12 April 2024).

Costello TH, Pennycook G and Rand D (2024) Durably reducing conspiracy beliefs through dialogues with AI. *OSF*. Available at: https://osf.io/xcwdn (accessed 12 April 2024).

Dobson JE (2023) On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities* 5(2): 431–449.

Grimmer J, Roberts ME and Stewart BM (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Kim A, Pethe C, Inoue N, et al. (2021) Cleaning dirty books: Post-OCR processing for previously scanned texts. In: *Findings of the association for computational linguistics: EMNLP 2021, November 2021, Punta Cana, Dominican Republic*, pp.4217–4226, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Latour B (1996) On interobjectivity. *Mind, Culture, and Activity* 3(4): 228–245.

Lupo L, Magnusson O, Hovy D, et al. (2023) How to use large language models for text coding: The case of fatherhood roles in public policy documents. *arXiv*:2311.11844 (accessed 12 April 2024).

Macanovic A and Przepiorka W (2024) A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from online posts. *Behavior Research Methods* 56: 2782–2803.

National Endowment for the Humanities (1861a) Arkansas true Democrat. [volume] (Little Rock, Ark.) 1857-1862, September 12, 1861, Image 2. 12 September. Available at: https://chroniclingamerica.loc.gov/lccn/sn82014282/1861-09-12/ed-1/seq-2/ (accessed 12 April 2024).

National Endowment for the Humanities (1861b) Macon beacon. [volume] (Macon, Miss.) 1859-1995, January 02, 1861, Image 2. 2 January. Available at: https://chroniclingamerica.loc.gov/lccn/sn83016943/1861-01-02/ed-1/seq-2/ (accessed 12 April 2024).

National Endowment for the Humanities (1861c) Memphis daily appeal. [volume] (Memphis, Tenn.) 1847-1886, November 21, 1861, Image 3. 21 November. Available at: https://chroniclingamerica.loc.gov/lccn/sn83045160/1861-11-21/ed-1/seq-3/ (accessed 12 April 2024).

National Endowment for the Humanities (1861d) Yorkville enquirer. [volume] (Yorkville, S.C.) 1855-2006, September 26, 1861, Image 2. 26 September. Available at: https://chroniclingamerica.loc.gov/lccn/sn84026925/1861-09-26/ed-1/seq-2/ (accessed 12 April 2024).

Navigli R, Conia S and Ross B (2023) Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality* 15(2): 10:1–10:21.

Neto AFS, Bezerra BLD and Toselli AH (2020) Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Applied Sciences* 10(21): 7711.

Ornstein JT, Blasingame EN and Truscott JS (2023) How to train your stochastic parrot: Large language models for political texts. Working Paper. Available at: https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf (accessed 1 August 2024).

Ouyang S, Zhang JM, Harman M, et al. (2023) LLM is like a box of chocolates: The non-determinism of ChatGPT in code generation. *arXiv*:2308.02828 (accessed 12 April 2024).

Rathje S, Mirea D-M, Sucholutsky I, et al. (2023) GPT is an effective tool for multilingual psychological text analysis. *OSF*. Available at: https://osf.io/sekf5 (accessed 12 April 2024).

Strobelt H, Webson A, Sanh V, et al (2023) Interactive and visual prompt engineering for Ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics* 29(1): 1146–1156.

Törnberg P (2023a) ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv*:2304.06588 (accessed 12 April 2024).

Törnberg P (2023b) How to use LLMs for text analysis. *arXiv*:2307.13106 (accessed 12 April 2024).

Xu Q, Wang L, Liu H, et al. (2022) LayoutLM-critic: Multimodal language model for text error correction of optical character recognition. In: Yang S and Lu H (eds) *Artificial Intelligence and Robotics*. Cham: Springer Nature, pp.136–146.

Ye W, Ou M, Li T, et al. (2023) Assessing hidden risks of LLMs: An empirical study on robustness, consistency, and credibility. *arXiv*:2305.10235 (accessed 12 April 2024).

Ziems C, Held W, Shaikh O, et al. (2024) Can large language models transform computational social science? *Computational Linguistics* 50(1): 237–291.

Author biography

Nicole Schwitter is a postdoctoral researcher at the Mannheim Centre for European Social Research and an honorary research fellow at the University of Warwick. Her research focuses on computational social science, social group dynamics, and issues related to discrimination and integration. In her research, she employs experimental methods and innovative approaches to answer social scientific questions.