DCPAPERS

Synthetic signal identification in LLM datasets

Jim Hahn[1,2]

[1]School of Information Sciences, University of Illinois, 501 E. Daniel St., Champaign, IL 61820-6211, USA
[2]Penn Libraries, University of Pennsylvania, 3420 Walnut Street, Philadelphia, PA 19104-6206, USA

Abstract

This research addresses the quality of training data in LLMs using methods from signaling theory and the talk page metadata of Wikipedia articles. The significance of the method is to lower the cost of information quality assessment in datasets. Natural language processing on metadata text generated sentiment, reading complexity, and self-reference scores as contributions to the computationally derived signals. Results showed that it is possible to understand indicators of information quality using textual computation over the metadata in article pages.

Keywords

large language models, dataset quality, synthetic signals, signaling theory

1. Introduction

Large Language Models (LLMs) like ChatGPT are profoundly influencing information retrieval and access. While there is much enthusiasm in utilizing ChatGPT-like services among metadata professionals and the public, studies have shown that LLMs provide false information [1, 2]. The false information may be a result of hallucination by the system [3]. Another cause may be that LLMs are trained on datasets of unknown quality [4, 5]. In one audit of over 18,000 datasets used to train AI models, encyclopedias were found to make up 21.5 percent of data sources, and Wikipedia is in 14.6 percent of the datasets [6]. The same analysis found that social media content (e.g., Reddit, Twitter, Quora) made up nearly 16 percent of the training data in AI systems.

The extent of problematic data found in public LLMs was documented in the paper "What's in my big data?" [4]. The authors found toxic content in training sets and content containing personally identifiable information. An attempt at retrospective documentation of the Colossal Clean Crawled Corpus advocated for three documentation levels of 1) metadata, 2) included data, and 3) excluded data [7]. The machine learning community is increasingly focusing on improving dataset discovery and reuse, as underscored by the recent introduction of the Croissant metadata format for datasets [8]. In the Hugging Face platform, dataset metadata is supporting improved documentation for some, though not all datasets. The aggregation "Dataset Cards with Metadata" (https://huggingface.co/datasets/librarian-bots/dataset_cards_with_metadata) is an automated dataset of metadata associated with all dataset cards in Hugging Face. According to a recent paper that analyzed the metadata associated with all Hugging Face platform dataset data

DCMI-2024 International Conference on Dublin Core and Metadata Applications
jimhahn@illinois.edu (J. Hahn)
0000-0001-7924-5294 (J. Hahn)

https://doi.org/10.23106/dcmi.952493684

cards, “the number of datasets on Hugging Face doubles every 18 weeks” [9]. As of March 14, 2024, there are 80,260 rows in the dataset cards with metadata dataset. However, the actual documentation for any dataset is limited as the same analysis noted, “Despite the importance of dataset cards, only 58.2 percent (14,011 out of 24,065 dataset repositories contributed by 4,782 distinct user accounts) include dataset cards as Markdown README.md files within their dataset repositories,” [9]. The notion of documenting excluded data, as noted in the retrospective analysis of the Colossal Clean Crawled Corpus is important as the exclusions may produce “representational harms” [7]. To summarize the findings from the retrospective documentation, the authors found “mentions of sexual orientations,” as having the “highest likelihood of being filtered out…” [7]. The Colossal Clean Crawled Corpus has the additional problem of containing benchmark data contamination and text in the corpus which was machine generated [7]. In the present research, the training data utilized for LLMs are the focus of applied signaling theory with the goals of improving information quality assessment.

1.1. Research Question

This research inquiry is focused on the question: “What computational methods can be devised to infer signals of true and untrue information from talk page metadata within Wikipedia artifacts?” There are two major outputs to this work: 1) a synthetic signals database that aggregated attributes of utterances from article talk pages in Wikipedia and 2) a method for development of a gold-standard utterance corpus that is thereafter used as a baseline comparison to Wikipedia articles of unknown quality.

2. Related Work

The Secure Learning Lab produced a “LLM Safety Leaderboard,” based on their DecodingTrust platform [10], “…organized around the following eight perspectives of trustworthiness: Toxicity, Stereotype and bias, Adversarial robustness, Out-of-distribution robustness, Privacy, Robustness to adversarial demonstrations, Machine ethics, fairness.”[1] To have a score approaching 100 is the highest rank, as the metrics for each area are from 0-100. A model must be submitted for scoring. Therefore, the ranking represents those models that have been submitted to the LLM safety leader board. The LLM Claude 2.0 is the top scoring LLM in the leaderboard. The dataset for Claude 2.0 is described in the model paper as “… a proprietary mix of publicly available information from the Internet, datasets that we license from third party businesses, and data that our users affirmatively share or that crowd workers provide” [11].

2.1. Signaling Theory

Signaling theory “models the relationship between signal and quality. For a signal to be reliable, the costs of deceptively producing the signal must outweigh the benefits. The core of signaling theory is its analysis of the types of signals and situations that bring this about” [12]. Costly signals are reliable, “because producing the signal requires possessing the indicated quality” [12].

[1] https://decodingtrust.github.io/

https://doi.org/10.23106/dcmi.952493684

Much like everyday language use, the signals in training data from the web are conventional signals. Conventional signals are difficult to assess and are unreliable sources of information. The costs to send and receive conventional signals are low [13].

Take for example, the costly signaling that a college degree indicates. This observation was the key contribution of Spence's seminal work in job market signals; underscoring for the purposes of this research another costly signal that is, in most cases, interpreted as a reliable signal [14]. The outlay of time and money it takes to pursue higher education in Spence's work was a reliable signal of a potential employee's value to an employer. Faking a degree is not simple to do in the long term, and to obtain a degree is thought to signal the quality of an employee.

Signaling theory is used as a departure point in this research and applied to information quality on the web. When dealing with information on the web, a space with many conventional signals, it is prohibitively expensive to the receiver to evaluate and identify signals that are reliable indicators of quality. This underscores the need for a method to ascertain quality of information quickly and easily. Signals can be derived by text analysis, as in the examples uncovered by Newman and others which found, "that liars tend to tell stories that are less complex, less self-relevant, and more characterized by negativity," [15]. The notion of negative statements as a cue to deception has been explored in related studies [16, 17]. Where conventional signals abound and there are not clear ways to detect a costly or reliable signal, researchers utilized account histories in social spaces (e.g. Twitter users) to generate synthesized signals that are derived in part from linguistic indicators [18]. The approach used in the present study is directly influenced from the work that mined account histories for synthesized signals [18].

3. Methods

This work first derived linguistic indicators for quality from the corpus of utterances and thereafter applied the indicators to two different sets of Wikipedia corpora. The first dataset explored was the gold-standard, drawn from "Good/Featured" Wikipedia article pages. The Wikipedia article categories of "Good/Featured" are selected by Wikipedia administrators and are characterized by "factual accuracy and verifiability."[2] This is contrasted to a Wikipedia article dataset of unknown quality.

3.1. Linguistic Indicators as Synthetic Signals

The mining for linguistic indicators was for three attributes that taken together could be the basis of synthetic signals identification: 1) the sentiment of the utterance, 2) the reading ease of the utterances, and 3) the presence of self-reference in the text. The NLP approaches include two different sentiment analysis tools. A third software library for NLP generated scoring for complexity of the utterances – a stand in for how difficult the text was to read. The first pass of sentiment mining used TextBlob.[3] A subsequent sentiment mining process then used Vader text mining software; these two scores were subsequently averaged [19]. The Flesch-

[2] https://w.wiki/8uUY

[3] https://textblob.readthedocs.io

https://doi.org/10.23106/dcmi.952493684

Kincaid readability score was generated over the text in talk pages to score reading level of the utterance.[4] The large sets of talk page data approach over 20 gigabytes of text for processing. A PySpark notebook was utilized to run the NLP tasks for creating, filtering, and combining Spark dataframes.[5] Spark is an alternative to MapReduce – employing a novel distributed processing for resilient distributed datasets [20]. Spark processing made this analysis possible, with a user defined function to make an additional attribute of self-reference by parsing the occurrences of "I" statements of each utterance. These correspond to Newman and others about untrue statements containing utterances which the speaker does not use self-reference [15, 16, 17].

4. Results

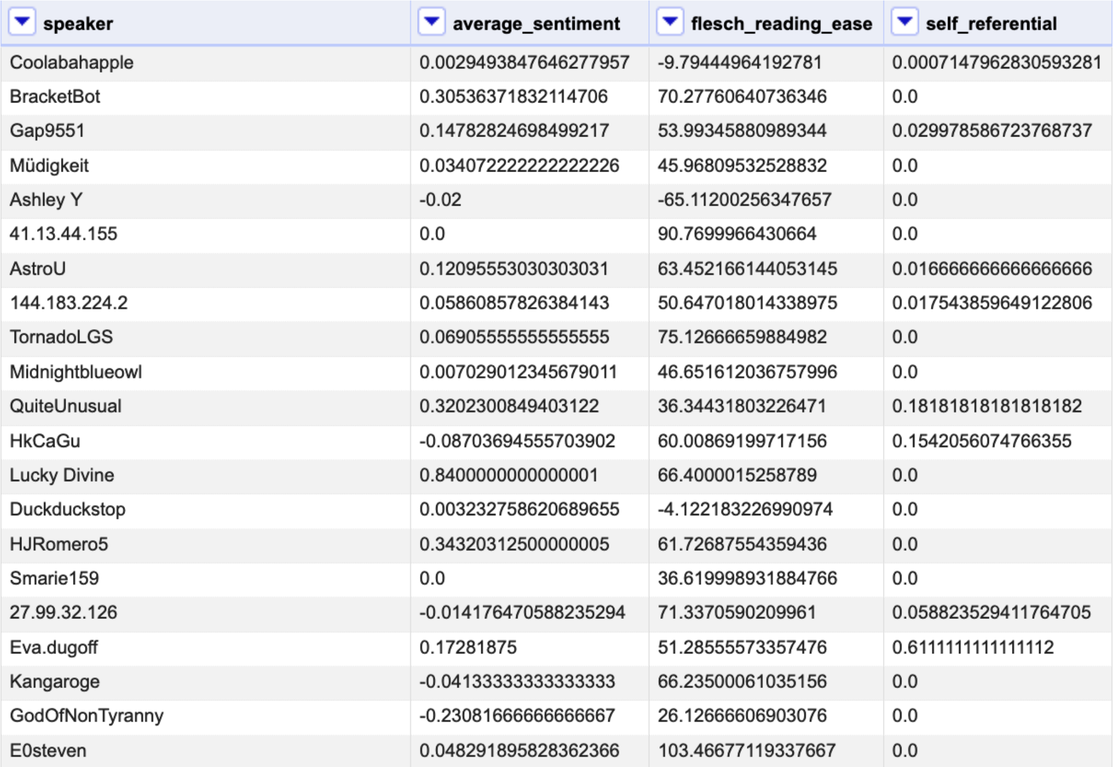
To evaluate the Wikipedia artifacts that correspond to high quality LLM datasets, the talk pages are first processed for the three attributes of sentiment, readability, and self-reference. To understand if computationally inferred synthetic signals of true and untrue information can be derived, a corresponding gold-standard of good quality article pages are needed. Two sentiment analysis tools are averaged together help to avoid a single point of failure for the NLP tasks of sentiment. For example, if one set of NLP software has a limitation for the given text, it may be possible to average out the limitation with another value of sentiment score. The conversations in the article talk pages [21] used for analysis were selected from 2016; this aligns with the gold-standard pages from the "Good/Featured" article dataset WikiText [22], also from 2016. The WikiConv dataset [21] contains talk page corpora from multiple years and were made accessible from the Python repository ConvoKit [23]. The Wikiconv dataset had a focus on understanding toxicity over all the years of talk pages published in the site. The dataset contains the years from 2001-2018, however, alignment with the 2016 talk pages was desirable for gold-standard computation. In 2016 there were 144,065 unique Wikipedia editors in the article talk page corpus. A result from the paper "Pointer Sentinel Mixture Models," was the curated set of Wikipedia article text from [22], which the authors made available as the Wikitext Corpus.[6] The Wikitext Corpus contains only those Wikipedia Articles pages from the "Good/Featured" pages in Wikipedia from the year 2016. This served as the high quality corpus. When the Wikitext corpus is paired alongside the article talk pages from 2016, a gold-standard of conversational text for quality pages can be derived. The Wikitext corpus contained 29,803 article pages.

4.1. Signal Database

The database of synthetic signals is organized first by all Wikipedia editors from the 2016 article talk pages [21], and then includes the aggregate average sentiment, average reading ease, and averages of the occurrences of self-referential statements for those editors. The database contains 144,065 rows. Of the 144,065 users, 22,692 (16 percent) have an average negative sentiment, 50,272 (35 percent) have an average positive sentiment, and 71,101 (49 percent) have an average neutral sentiment. Figure 1 shows values for indicators utilized in synthetic

[4] https://github.com/shivam5992/textstat
[5] https://pypi.org/project/pyspark/
[6] https://huggingface.co/datasets/wikitext

https://doi.org/10.23106/dcmi.952493684



speaker	average_sentiment	flesch_reading_ease	self_referential
Coolabahapple	0.0029493847646277957	-9.79444964192781	0.0007147962830593281
BracketBot	0.30536371832114706	70.27760640736346	0.0
Gap9551	0.14782824698499217	53.99345880989344	0.029978586723768737
Müdigkeit	0.034072222222222226	45.96809532528832	0.0
Ashley Y	-0.02	-65.11200256347657	0.0
41.13.44.155	0.0	90.7699966430664	0.0
AstroU	0.12095553030303031	63.452166144053145	0.016666666666666666
144.183.224.2	0.05860857826384143	50.647018014338975	0.017543859649122806
TornadoLGS	0.06905555555555555	75.12666659884982	0.0
Midnightblueowl	0.007029012345679011	46.651612036757996	0.0
QuiteUnusual	0.3202300849403122	36.34431803226471	0.18181818181818182
HkCaGu	-0.08703694555703902	60.00869199717156	0.1542056074766355
Lucky Divine	0.8400000000000001	66.4000015258789	0.0
Duckduckstop	0.003232758620689655	-4.122183226990974	0.0
HJRomero5	0.34320312500000005	61.72687554359436	0.0
Smarie159	0.0	36.619998931884766	0.0
27.99.32.126	-0.014176470588235294	71.3370590209961	0.058823529411764705
Eva.dugoff	0.17281875	51.28555573357476	0.6111111111111112
Kangaroge	-0.04133333333333333	66.23500061035156	0.0
GodOfNonTyranny	-0.23081666666666667	26.12666606903076	0.0
E0steven	0.048291895828362366	103.46677119337667	0.0

Figure 1: Selected Values from the Synthetic Signal Database

signal development. The sentiment of utterances is a conventional sentiment score from -1 to 1. Thereafter, the plots that follow in this paper were delineated in increments of .25. Using the signal database joined with editors from the "Good/Featured" quality pages, a resulting analysis in Figure 2, showed findings for sentiment. Figure 2 shows that for a good quality page, the discussion need not be completely positive. The gold-standard data primarily includes neutral to somewhat negative utterances. Good quality pages are not characterized by extremes (e.g. not very positive or very negative). The gold-standard dataset does not include editor text that have low complexity, no self-reference, and are very negative. The unknown quality Wikipedia pages contain more extremes in sentiment, shown in Figure 3. Specifically, the extremes which are observed in unknown quality Wikipedia article talk pages include more utterances of highly positive sentiment.

5. Discussion

The database of synthetic signals paired with a known quality set of articles helped to characterize the conversations that contribute to quality information. This was an expected result. However, the nature of the sentiment found, trending to somewhat neutral and even in the range of somewhat negative as interpreted by sentiment mining tools was not expected. It may

https://doi.org/10.23106/dcmi.952493684

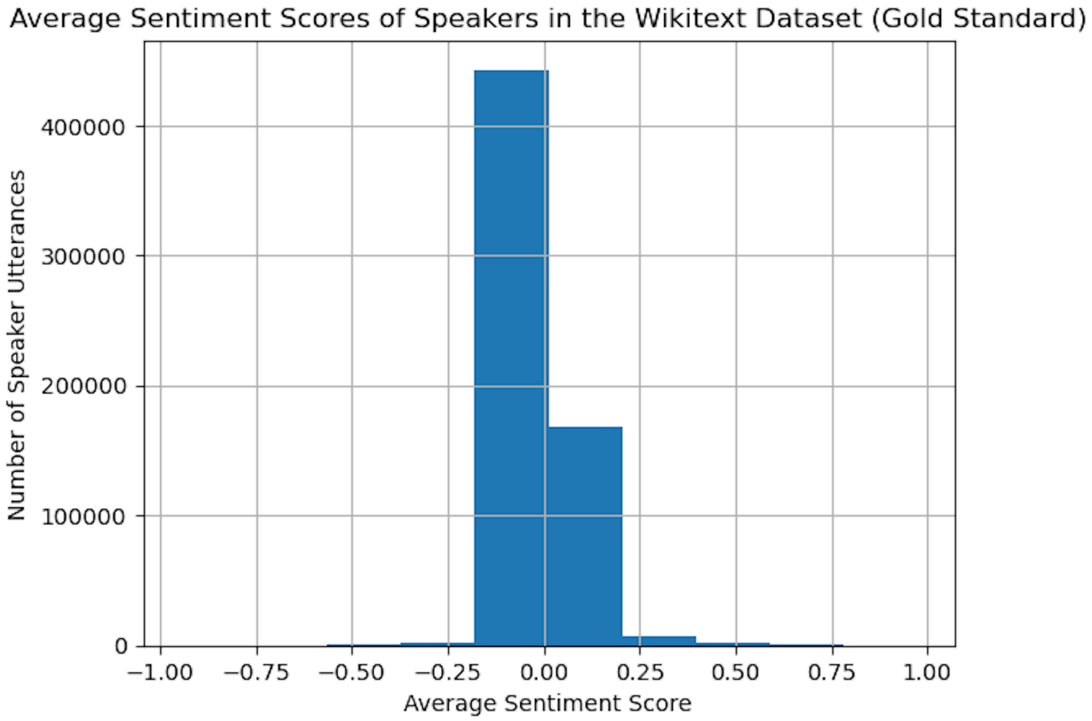


Figure 2: Sentiment of the Gold-Standard Corpus

be the case that to create good quality articles there is a need for conversation sentiment that is not characterized by extremes. Almost seventy percent of the conversational sentiment of the gold-standard was between -0.179 and 0.128. The meaning of this is not clear, though it may point to the qualitative nature of speech that characterizes critical thinking. Critical thinking is the process of inquiring about what is known and how it is known, it asks what evidence supports a given claim.

5.1. Social Spaces and Collaboration

Another interpretation for the Wikipedia talk page sentiment that characterizes the gold-standard set, is an interpretive perspective that takes into account the asynchronous nature of the social space. The interactions of Wikipedia talk pages were designed for collaboration. Wikipedia talk pages are not designed for real-time conversations. A Wikipedia editor participating in a talk page discussion does not need to respond immediately and can take time to find resources and re-read statements and texts, before replying. From a cognitive standpoint this allows the participants in the conversation to shift into what Kahneman described as system 2 reasoning, characterized by engaging logic and deeper thought [24]. Concerning speed of interaction, in a study that contrasts synchronous to asynchronous communication, email was shown to be the most truthful of internet-based communication technologies [25]. When

https://doi.org/10.23106/dcmi.952493684

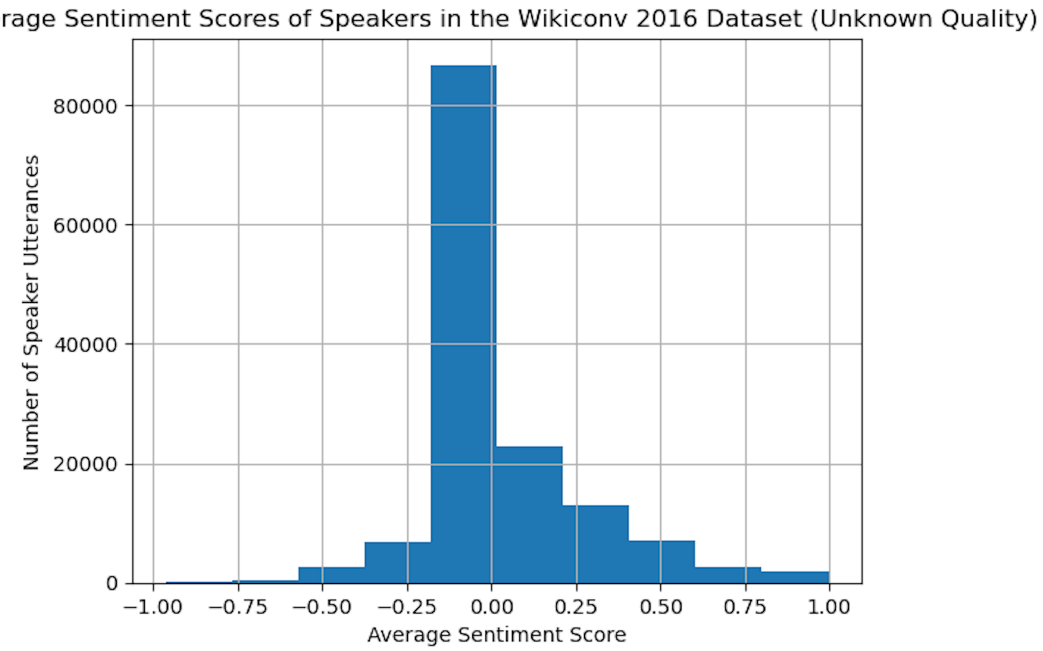


Figure 3: Sentiment of the Unknown Quality Corpus

viewed in the context of talk pages it is possible to infer that the talk pages might be a place of more authentic speech when given the opportunity for discourse to take place asynchronously [25]. This stands in contrast to other social media where reactions and speech elevate quickly because of the designed mobile notifications. This observation underscores the need to more critically evaluate how datasets are chosen for LLM training. While there may be an ample supply of text to mine on the web, some areas of the web that generate text, such as the spaces with rapid notifications and designed for high engagement require additional curating and selection of true or authentic speech by applying additional computation, such as the synthetic signals derived in the present study and in studies of social networking accounts [18], which may help to exclude from LLM training sets those data that are either inauthentic, toxic, or that propagate misinformation.

6. Conclusions

In general, discovering synthetic signals [18] could help to identify quality of information in Wikipedia datasets of unknown quality. Computationally identifying indicators of dataset quality from a linguistic standpoint [15] may guide LLM developers in selecting more trustworthy data for training and refining of LLMs. The aim of this method is to improve quality of LLMs by understanding reliable signals of trustworthy data that are too costly to fake, and low cost to the receivers [18]. The cost to fake a signal would be the development of a history of editor actions

https://doi.org/10.23106/dcmi.952493684

which contain low negativity, more self-reference, and a higher complexity of writing. The methods in this paper were focused on English language indicators. However, one of the great advantages of Wikipedia is the broad support for world languages. Related work inspected Polish language and untrue statements and did not find differences in spoken and written statements for their truth value [26], which is a contrast to findings from English language written and spoken modalities with text from email found to be more authentic when contrasted to spoken communication [25]. Detection of authentic speech is challenging for non-native language speakers and found they "appear to be at a significant disadvantage in lie-detection contexts" [27]. These examples underscore the complexity in ascertaining or evaluating authenticity in cultural contexts outside of native English-speaking spaces on the web. To expand this approach to other information sources on the web outside of Wikipedia, a process to abstract the discussion about information from the information itself must be identified. When such an abstraction can be modeled then the NLP processes about the information can be systematically mined for synthetic signals. The pilot method developed here showed promise for developing synthetic signals in Wikipedia artifacts.

Acknowledgments

My sincere thanks to Prof. Karrie Karahalios for teaching Social Spaces on the Internet, Fall 2023 at the University of Illinois. This work is a direct output from course readings, lectures, and my semester project.

References

[1] B. Mittelstadt, S. Wachter, C. Russell, To protect science, we must use LLMs as zero-shot translators, Nature Human Behaviour 7 (2023) 1830–1832. URL: https://www.nature.com/articles/s41562-023-01744-0. doi:10.1038/s41562-023-01744-0.

[2] G. P. Reddy, Y. V. Pavan Kumar, K. P. Prakash, Hallucinations in large language models (LLMs), in: 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2024, pp. 1–6. doi:10.1109/eStream61684.2024.10542617.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.

[4] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, J. Dodge, What's in my big data?, 2023. URL: http://arxiv.org/abs/2310.20707. doi:10.48550/arXiv.2310.20707, arXiv:2310.20707.

[5] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, M. De Choudhury, Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–20. URL: https://doi.org/10.1145/3544548.3581318. doi:10.1145/3544548.3581318.

[6] S. Longpre, R. Mahari, A. Chen, N. Obeng-Marnu, D. Sileo, W. Brannon, N. Muennighoff,

N. Khazam, J. Kabbara, K. Perisetla, X. Wu, E. Shippole, K. Bollacker, T. Wu, L. Villa, S. Pentland, S. Hooker, The data provenance initiative: A large scale audit of dataset licensing & attribution in AI, 2023. URL: http://arxiv.org/abs/2310.16787, arXiv:2310.16787.

[7] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting large webtext corpora: A case study on the colossal clean crawled corpus, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1286–1305. URL: https://aclanthology.org/2021.emnlp-main.98. doi:10.18653/v1/2021.emnlp-main.98.

[8] M. Akhtar, O. Benjelloun, C. Conforti, J. Giner-Miguelez, N. Jain, M. Kuchnik, Q. Lhoest, P. Marcenac, M. Maskey, P. Mattson, L. Oala, P. Ruyssen, R. Shinde, E. Simperl, G. Thomas, S. Tykhonov, J. Vanschoren, S. Vogler, C.-J. Wu, Croissant: A metadata format for ML-ready datasets, 2024. arXiv:2403.19546.

[9] X. Yang, W. Liang, J. Zou, Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face, 2024. arXiv:2401.13822.

[10] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models, 2024. arXiv:2306.11698.

[11] Anthropic, Model Card and Evaluations for Claude Models, 2024. URL: https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf, [Accessed 17-03-2024].

[12] J. Donath, Signals in Social Supernets, Journal of Computer-Mediated Communication 13 (2007) 231–251. URL: https://doi.org/10.1111/j.1083-6101.2007.00394.x. doi:10.1111/j.1083-6101.2007.00394.x.

[13] J. Donath, Identity and deception in the virtual community, in: Communities in Cyberspace, Routledge, London, 1998, pp. 27–57. URL: https://smg.media.mit.edu/papers/Donath/IdentityDeception/IdentityDeception.pdf.

[14] M. Spence, Job Market Signaling, The Quarterly Journal of Economics 87 (1973) 355–374. URL: https://academic.oup.com/qje/article-lookup/doi/10.2307/1882010. doi:10.2307/1882010.

[15] M. L. Newman, J. W. Pennebaker, D. S. Berry, J. M. Richards, Lying Words: Predicting Deception from Linguistic Styles, Personality and Social Psychology Bulletin 29 (2003) 665–675. URL: http://journals.sagepub.com/doi/10.1177/0146167203029005010. doi:10.1177/0146167203029005010.

[16] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, H. Cooper, Cues to deception, Psychological Bulletin 129 (2003) 74–118. doi:10.1037/0033-2909.129.1.74.

[17] F. Abri, L. F. Gutierrez, A. S. Namin, K. S. Jones, D. R. W. Sears, Linguistic features for detecting fake reviews, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Miami, FL, USA, 2020, pp. 352–359. URL: https://ieeexplore.ieee.org/document/9356253/. doi:10.1109/ICMLA51294.2020.00063.

[18] J. Im, S. Tandon, E. Chandrasekharan, T. Denby, E. Gilbert, Synthesized social signals: Computationally-derived social signals from account histories, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for

Computing Machinery, New York, NY, USA, 2020, p. 1–12. URL: https://doi.org/10.1145/3313831.3376383. doi:10.1145/3313831.3376383.

[19] C. Hutto, E. Gilbert, VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. doi:10.1609/icwsm.v8i1.14550.

[20] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, I. Stoica, Resilient distributed datasets: A Fault-Tolerant abstraction for In-Memory cluster computing, in: 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), USENIX Association, San Jose, CA, 2012, pp. 15–28. URL: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia.

[21] Y. Hua, C. Danescu-Niculescu-Mizil, D. Taraborelli, N. Thain, J. Sorensen, L. Dixon, WikiConv: A corpus of the complete conversational history of a large online collaborative community, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2818–2823. URL: http://aclweb.org/anthology/D18-1305. doi:10.18653/v1/D18-1305.

[22] S. Merity, C. Xiong, J. Bradbury, R. Socher, Pointer sentinel mixture models, 2016. arXiv:1609.07843.

[23] J. P. Chang, C. Chiam, L. Fu, A. Wang, J. Zhang, C. Danescu-Niculescu-Mizil, ConvoKit: A toolkit for the analysis of conversations, in: O. Pietquin, S. Muresan, V. Chen, C. Kennington, D. Vandyke, N. Dethlefs, K. Inoue, E. Ekstedt, S. Ultes (Eds.), Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 1st virtual meeting, 2020, pp. 57–60. URL: https://aclanthology.org/2020.sigdial-1.8. doi:10.18653/v1/2020.sigdial-1.8.

[24] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux, New York, 2011.

[25] J. T. Hancock, J. Thom-Santelli, T. Ritchie, Deception and design: The impact of communication technology on lying behavior, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Vienna Austria, 2004, pp. 129–134. URL: https://dl.acm.org/doi/10.1145/985692.985709. doi:10.1145/985692.985709.

[26] J. Sarzynska-Wawer, A. Pawlak, J. Szymanowska, K. Hanusz, A. Wawer, Truth or lie: Exploring the language of deception, PLOS ONE 18 (2023) e0281179. URL: https://dx.plos.org/10.1371/journal.pone.0281179. doi:10.1371/journal.pone.0281179.

[27] E. Elliott, A.-M. Leach, False impressions? The effect of language proficiency on cues, perceptions, and lie detection., Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement 56 (2024) 31–40. doi:10.1037/cbs0000337.