Measuring Inclusion through Multimodal Large Language Models

Armanda Lewis[1,*,†]

[1]New York University, 100 Washington Square East, New York, NY, 10003, USA

Abstract

Advancing inclusion within groups is important given the concept's direct links to improved business and scientific collaborations, sustained group learning, and enhanced creativity. This presentation describes an ongoing project on how we may use multimodal large language models to detect and measure inclusion as a set of multimodal (text, audio, and video) behaviors and linguistic markers. The main deliverables of this project are multimodal LLMs finetuned to detect inclusion and the prototyping of a feedback system that will help people learn to be more inclusive. Our models are informed by human-created inclusion labels. Large language models of varying size and modality are explored, including BERT-base, larger unimodal language models, and multimodal language models. Broadly, this project advances our knowledge of multimodality within natural language processing (NLP), informs how language models may be used ethically in downstream tasks, and promotes inclusion within society.

Keywords

multimodal large language models, inclusion, human annotation, NLP

1. Introduction

This research project, which works towards an automated inclusion detection system, sits at the intersection of natural language processing (NLP), the learning sciences, and human annotation. We describe ongoing efforts to develop multimodal large language models (LLMs) - massive machine learning models equipped to detect and generate language and incorporate non-verbal inputs - in order to detect and measure inclusion within small groups. There exists evidence to show that inclusion, the degree to which individuals experience treatment from a group that satisfies their need for belongingness and uniqueness [1], is linked to improved scientific and business collaborations [2, 3], sustained learning and adaptive change within diverse groups [4], and improved creativity [5]. LLMs are finetuned to detect inclusive behaviors and language, and then used to generate text-based user feedback to promote inclusive practices among small groups.

Positive impacts on inclusion have been gleaned from research that primarily relies on individual survey responses and qualitative observations. Since automatically and continuously

AIED2023: Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, July 07, 2023, Tokyo, Japan

*Corresponding author.

al861@nyu.edu (A. Lewis)

0000-0001-6189-494X (A. Lewis)

Dimensions	Low belongingness	High belongingness
Low uniqueness	Exclusion	Assimilation
High uniqueness	Differentiation	Inclusion

Table 1
Inclusion Framework [1]

measuring the psychological experience of individuals is not feasible, we focus on operationalizing inclusion as a set of observable, multimodal (text, audio, and video) behaviors, and linguistic markers that form features of finetuned large language models. LLMs have potential for modeling inclusion given their ability to both model and generate language [6], though their full capabilities have yet to be fully understood [7]. To the author's knowledge, this study is the first of its kind to explore inclusion through a multimodal lens and through large, generative NLP models. This project aims to answer: How may we use multimodal large language models to measure inclusion quality automatically within specific small group interactions, and how may we use those models to generate simple feedback that supports groups to engage more inclusively?

2. Ongoing Work

Inspired by work on collaboration detection [8, 9], we first mapped the theoretical inclusion dimensions listed in Table 1 to observable behaviors and language. Observable information is then translated to model features, which will provide us with operationalized inclusion dimensions grounded in theory. The project utilizes traditional NLP modeling methods [10, 11], in which experiments are conducted using pre-existing language models that were trained on enormous amounts of text and dialogue data from the internet and digitized texts (i.e. pretrained models). The author expores new methods to have pretrained models finetuned on multimodal data, an emerging area of large language model research [12, 13, 14]. As such, we utilize several pretrained large language models that increase in scale, including BERT-base [15], GPT-2 [16], and GPT-3 (GPT-J, BLOOM, and GPT 3.5) [17, 18], as well as in modality. Models are finetuned on unimodal and multimodal group interaction data (audio, text, and/or video frames). Learning rates and batch sizes will be determined according to standard task settings, and follow the training-test splits and standards articulated by [19] and [20]. A supervised baseline classification paradigm is adopted to predict inclusion labels that follow the inclusion dimensions, and increasingly scaled models are tested to examine the expressive power of our models.

In order to develop human-centered LLMs, modeling will be grounded by a crowdsourcing experiment, where general skill crowdworkers and inclusion experts manually label multimedia snippets according to the inclusion framework. In addition to exploring both feature-specific and end-to-end approaches [21], the project will conduct an experiment of removing content-based task language to see how well the model can detect inclusion. Model performance is compared in terms of classification accuracy of our expert and general classification scores on inclusion dimensions. Comparing different unimodal and multimodal model performance will serve as an ablation approach to examine the role of feature modalities in terms of overall model performance. The best performing model will be used within a small proof-of-concept experiment where inclusive behaviors are detected within small group interactions, and simple text-based feedback is generated based on the exchanges.

3. Data Collection Process

We will also detail two important data collection aspects of the work, which will relate to robust use of large language models in downstream tasks. One is the creation of a multimodal group interaction training dataset compiled from open source datasets such as the AMI corpus [22], which contains audio and video of cooperative problem solving scenarios, and the newly released CANDOR corpus [23], a large multimodal dataset of naturalistic conversations. Our finetuning process will incorporate these multimodal data, as well as open source NLP datasets developed for linguistic bias detection and model evaluation [24, 25]. This compiled corpus also will be used for crowdworker annotations, described below.

One priority of this project is to explore human-grounded models of inclusion. To this end, and following the general dataset collection procedures described in [26], we will gather human annotations based on our inclusion indicators of random samples taken from our compiled corpus. Annotators include crowdworkers recruited from Amazon Mechanical Turk, and advanced graduate students with expertise in inclusion theory. The expert data will form ground truth labels, and manual and statistical analyses will reveal any divergences between expert and general skill annotators. There will also be a validation phase, where we will gauge interrater agreement between annotators of each group.

The crowdsourced annotation experiment to collect human labels of inclusive exchanges will be approved by our Institutional Review Board to ensure integrity of procedures and maintenance of ethical and transparent practices for recruited participants. We will collect expert annotations that will then inform guidance for generally skilled Amazon Mechanical Turk workers, using NLP crowdsourcing research insights to ensure a fair payment structure and ethical treatment [27, 28, 29]. We will store all data on secure servers and all data will be processed and managed on a secure local machine (e.g., lockable computer systems with passwords, firewall system in place, power surge protection, virus/malicious intruder protection), or a secure institutional high performance computing environment. This study will collect non-sensitive data, and we adhere to any licensing specifics of our open source corpora. We intend on releasing our human labels and open access data snippets as an open source dataset that can be used by computational scientists to gauge human-model alignment issues that emerge in the downstream use of advanced machine learning models and by social scientists to explore how humans view inclusion and its dimensions.

4. Presentation of Ongoing Work

This presentation will detail results of human annotation, and divergences between annotators, and between human and model characterizations. We will discuss initial efforts in creating multimodal large language models for downstream tasks, and will also detail the importance of grounding large language model resarch within robust theories.

References

[1] L. M. Shore, A. E. Randel, B. G. Chung, M. A. Dean, K. Holcombe Ehrhart, G. Singh, Inclusion and Diversity in Work Groups: A Review and Model for Future Research, Journal of Management 37 (2011) 1262–1289.

[2] M. E. Mor Barak, The Inclusive Workplace: An Ecosystems Approach to Diversity Management, Social Work 45 (2000) 339–353.

[3] I. M. Nembhard, A. C. Edmondson, Making it safe: the effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams, Journal of Organizational Behavior 27 (2006) 941–966.

[4] R. J. Ely, D. A. Thomas, Cultural Diversity at Work: The Effects of Diversity Perspectives on Work Group Processes and Outcomes, Administrative Science Quarterly 46 (2001) 229–273.

[5] H. Leroy, C. Buengeler, M. Veestraeten, M. Shemla, I. J. Hoever, Fostering Team Creativity Through Team-Focused Inclusion: The Role of Leader Harvesting the Benefits of Diversity and Cultivating Value-In-Diversity Beliefs, Group & Organization Management 47 (2022) 798–839.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, arXiv:2005.14165 [cs] (2020). URL: http://arxiv.org/abs/2005.14165, arXiv: 2005.14165.

[7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Kohd, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs] (2021). URL: http://arxiv.org/abs/2108.07258, arXiv: 2108.07258.

[8] S. Praharaj, M. Scheffel, M. Schmitz, M. Specht, H. Drachsler, Towards Automatic Collaboration Analytics for Group Speech Data Using Learning Analytics, Sensors 21 (2021) 3156. URL: https://www.mdpi.com/1424-8220/21/9/3156. doi:10.3390/s21093156.

[9] M. Worsley, X. Ochoa, Towards collaboration literacy development through multimodal learning analytics, in: Companion Proceedings 10th International Conference on Learning

Analytics & Knowledge (LAK20), volume 2610, 2020, pp. 53–63. URL: http://ceur-ws.org/Vol-2610/paper11.pdf.
[10] D. Jurafsky, J. H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall series in artificial intelligence, 2nd ed ed., Pearson Prentice Hall, Upper Saddle River, N.J, 2009.
[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing, arXiv:1910.03771 [cs] (2020). URL: http://arxiv.org/abs/1910.03771, arXiv: 1910.03771.
[12] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, 2023.
[13] M. K. Hasan, M. S. Islam, S. Lee, W. Rahman, I. Naim, M. I. Khan, E. Hoque, TextMI: Textualize Multimodal Information for Integrating Non-verbal Cues in Pre-trained Language Models, 2023.
[14] S. Janghorbani, G. de Melo, MultiModal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision Language Models, 2023.
[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (2018).
[16] A. Radford, K. Narasimhan, Improving Language Understanding by Generative Pre-Training, undefined (2018). URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.
[17] B. Wang, A. Komatsuzaki, GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, https://github.com/kingoflolz/mesh-transformer-jax, 2021.
[18] B. Workshop, BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2023.
[19] M. Guo, Y. Yang, D. Cer, Q. Shen, N. Constant, MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models, arXiv:2005.02507 [cs] (2020). URL: http://arxiv.org/abs/2005.02507, arXiv: 2005.02507.
[20] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep Learning Based Text Classification: A Comprehensive Review, arXiv:2004.03705 [cs, stat] (2021). URL: http://arxiv.org/abs/2004.03705, arXiv: 2004.03705.
[21] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, S. Young, A Network-based End-to-End Trainable Task-oriented Dialogue System, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 438–449.
[22] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, P. Wellner, The AMI Meeting Corpus: A Pre-announcement, in: S. Renals, S. Bengio (Eds.), Machine Learning for Multimodal Interaction, volume 3869, Springer

Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 28–39. URL: http://link.springer.com/10.1007/11677482_3. doi:10.1007/11677482_3.
[23] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, S. Marin, The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation, Science Advances 9 (2023) eadf3197.
[24] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, A. Williams, Multi-Dimensional Gender Bias Classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 314–331.
[25] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social Bias Frames: Reasoning about Social and Power Implications of Language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490.
[26] H. He, A. Balakrishnan, M. Eric, P. Liang, Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings, arXiv:1704.07130 [cs] (2017). URL: http://arxiv.org/abs/1704.07130, arXiv: 1704.07130.
[27] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv:1508.05326 [cs] (2015). URL: http://arxiv.org/abs/1508.05326, arXiv: 1508.05326.
[28] N. Nangia, S. Sugawara, H. Trivedi, A. Warstadt, C. Vania, S. R. Bowman, What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1221–1235.
[29] B. Shmueli, J. Fell, S. Ray, L.-W. Ku, Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3758–3769.