SAIS 2024 Proceedings

Southern (SAIS)

Spring 3-16-2024

LLMs For Gender Prediction: A Comparative Study

Samira Rahman

Md Jabir Rahman

LLMS FOR GENDER PREDICTION: A COMPARATIVE STUDY

Extended Abstract

Samira Rahman
The University of Memphis
srahman1@okcu.edu

Md Jabir Rahman
Oklahoma City University
mrahman@okcu.edu

ABSTRACT

Given the elevated popularity of Large Language Models (LLMs) and their accessible APIs for analyzing large-scale datasets, users and researchers are increasingly likely to leverage these models for data labeling tasks. Although LLMs hold immense promise for streamlining the costly and labor-intensive data labeling process, more research is needed to explore and compare their performance against human coders. This study addresses this gap by employing a suite of models, including a Python-based deep learning package and various state-of-the-art LLMs, to assess their alignment with human annotations. Our findings uncover significant insights into the capabilities and limitations of LLMs, offering valuable contributions to the academic and industrial spheres.

Keywords

ChatGPT-4, ChatGPT-3, Gender-Guesser, Google Gemini, LLM Performance

EXTENDED ABSTRACT

Understanding gender dynamics is crucial in online peer-to-peer (P2P) business models, where gender can significantly influence trust, interaction, and transaction decisions (Edelman & Luca, 2014; Marchenko, 2019). For businesses, accurate gender identification can provide insights into consumer behavior, aid in personalizing user experiences, and help mitigate biases, making it a pertinent study area (Cui et al., 2020). In digital contexts, automating accurate gender identification is essential as automation can be helpful, especially when handling large data sets where users must use their names for the service (e.g., Airbnb, Turo), and manual annotation is less practical and/or costly. With the emergence of Generative Pre-trained Transformers such as ChatGPT and Google's Gemini, there is a growing interest in utilizing GPTs for data labeling. Our research thoroughly examines this trend by highlighting the capabilities of Large Language Models (LLMs) and deep learning models. It further cautions researchers against the uncritical acceptance of GPT outputs without subsequent validation.

In this study, we evaluate the performance of three artificial intelligence models—GPT-3, GPT-4, and a Python-based gender guesser (gender-guesser 0.4.0)—in gender identification from first names, comparing their outcomes to human-labeled benchmarks. Out of an initial set of 500 true Airbnb hosts' first names, 484 were selected for analysis, with the rest excluded due to ambiguous gender associations. Human references were established through comprehensive Google image searches for ambiguous names to determine each name's predominant gender representation. Additionally, we apply (Google's Gemini) on randomly selected one hundred names from our initial set for a comparison with a state-of-the-art model.

Using Hosts' first names as inputs, Python-based gender guesser leads with an accuracy of 98.76%, closely followed by GPT-3 at 98.14%, while GPT-4 surprisingly lags at 78.10%. The Python guesser and GPT-3 share a high label similarity rate of 98.97%, suggesting a robust alignment. Conversely, both demonstrate significantly lower congruence with GPT-4, highlighting a disparity in their gender recognition patterns. The investigation also pinpointed specific names mislabeled by each AI model, shedding light on the diverse challenges inherent in AI-driven gender identification. While Gemini showed an impressive 99% accuracy, it lacked the flexibility to feed a large dataset, and we needed to input each name separately.

The study underscores AI's potential and current limitations in large scale gender identification, possibly influenced by cultural, linguistic, and contextual elements. While the high accuracy of the Python guesser and GPT-3 and Gemini is promising for academic research and enhancing user experience in P2P platforms, the discrepancies and lower performance of GPT-4 emphasize the need for continued enhancement of AI models and caution while using those. We intend to broaden this investigation by incorporating an expanded dataset and augmenting it with more extensive annotations from online platforms such as Amazon MTurk or Prolific. Subsequent studies should concentrate on enhancing the precision of models and meticulously contrasting various algorithms specifically for tasks such as automated gender determination. This research endeavors to deepen the comprehension of Large Language Models and task automation, contributing significantly to the scholarly discourse in IS literature.

REFERENCES

1. Cui, R., Li, J., & Zhang, D. J. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science*, 66(3), 1071-1094.Edelman, B. G., &
2. Luca, M. (2014). Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054).
3. Marchenko, A. (2019). The impact of host race and gender on prices on Airbnb. *Journal of Housing Economics*, 46, 101635.