Alessia Pensabene

Scuola Normale Superiore

How real is synthetic data?

Memes at work

Team Members

Facilitators: Margherita Di Cicco, Richard Rogers, Asaf Nissenbaum.
Participants: Elia Meregalli, Clara Eck, Zuzana L’udvikova, Xinmeng Hu, Ling Tuns Tsang, Krisztina Vizoli, Zeyu Yang, Yiran Hu, Shiqi Fu, Yutong Ma, Xinyi Chen, Yu Chenm Yunyi Zhang (Echo), Zhiyao Luo (Leo), Xiaoyun Huang, Lin Shi, Ruiwen Zhou, Nuo Chen, Alessia Pensabene.
Designers: Elena Aversa, Alessandra Facchin.

Contents

1. Introduction

The rise of artificial intelligence and machine learning has led to the increasing use of synthetic data—artificially generated information designed to mimic real-world data—for research and analysis. While synthetic data is often used in fields like data privacy, predictive modeling, and AI training (Rubin, 1993; Jordon et al., 2018), its reliability in accurately representing real-world patterns remains an open question (Birhane et al., 2021).

In this study, we explore the validity of synthetic data by examining related hashtags generated by large language models (LLMs) like ChatGPT and comparing them to actual Instagram data. Specifically, we investigate whether the related hashtags suggested by the model appear in real Instagram posts and whether they co-occur as frequently as the model predicts. Understanding the accuracy and limitations of synthetic data is essential for determining its utility in social media research and broader data-driven applications (Hao, 2021).

Prior research has shown that LLMs generate outputs based on statistical correlations in their training data rather than direct real-world observations (Bender et al., 2021). As a result, their ability to generate reliable social media data is questionable, given that they lack real-time access to Instagram's ever-evolving digital landscape. By analyzing both the presence and relationships of hashtags in synthetic and real-world data, we aim to assess whether LLM-generated information aligns with real-world digital objects and to what extent it can serve as a viable tool for social media analysis. This study contributes to ongoing discussions about the role of AI in data generation, highlighting both its potential and its limitations in accurately reflecting real-world trends.

2. Initial Data Sets

Our study utilizes two main data sets:

- **Synthetic Data**: This data set was generated by prompting ChatGPT to suggest related hashtags based on a set of initial hashtags. For example, we provided hashtags such as #work, #quietquitting, #quietquittinglife, #antiworkmemes and asked ChatGPT to generate additional related hashtags. Specifically, we requested five hashtags associated with each of the initial hashtags, ensuring that they were relevant for studying the social imaginaries created through work-related memes on Instagram. We conducted multiple experiments, modifying our prompts to explore different ways in which the model determines hashtag relationships. This approach allowed us to analyze whether the model maintained consistency across different queries and how it established connections between hashtags.

- **Real-World Data**: This data set consists of hashtags and their associated posts, retrieved both manually and through automated software. The posts were analyzed to determine the actual occurrence and co-occurrence of the hashtags provided by ChatGPT.

Both data sets were compared to assess the extent to which the synthetic hashtags generated by ChatGPT align with real-world hashtag usage patterns on Instagram.

3. Research Questions

1. How accurately do related hashtags generated by LLMs reflect their presence on Instagram?
2. What criteria do LLMs use to determine "relatedness" among hashtags?
3. To what extent does the analysis of synthetic data align with real-world data?
4. How productive is synthetic data for analytical purposes compared to manual analysis?

4. Methodology

To systematically compare synthetic and real-world data, we employed a combination of automated tools and manual methods for data collection, processing, and analysis. Our approach consisted of three main steps: synthetic hashtag generation, real-world data collection, and comparative analysis.

1. **Synthetic Hashtag Generation**

We used ChatGPT to generate lists of related hashtags for specified topics. The process involved:

- Selecting initial seed hashtags related to work and labor discourse on social media, such as #work, #quietquitting, #quietquittinglife, #antiworkmemes.
- Asking ChatGPT to provide five related hashtags for each seed hashtag.
- Requesting ChatGPT to explain the rationale behind the relationships, analyzing whether the connections were based on semantic similarity, thematic relevance, or co-occurrence patterns inferred from training data.

- Conducting iterative prompting experiments, adjusting our queries to explore variations in the model's responses and its consistency in suggesting related hashtags.

2. **Real-World Data Collection**

To assess the presence and relationships of the ChatGPT-generated hashtags in real Instagram posts, we employed a combination of manual and automated methods:

- Manual collection: We searched for Instagram posts containing some hashtags and examined their associated hashtags to check whether ChatGPT's suggested hashtags appeared in real usage.

- We then compiled a dataset containing real hashtags.

3. Comparative Analysis

Once both synthetic and real-world data sets were compiled, we performed a structured comparison using multiple techniques:

- Co-hashtag graphs: We visualized hashtag relationships by constructing network graphs, mapping connections between the original hashtags and the ChatGPT-generated hashtags based on Instagram occurrences.
- Manual coding and labeling: We categorized hashtags into thematic clusters manually to evaluate conceptual alignment with ChatGPT's clusters.
- Automated clustering: We also used ChatGPT to categorize hashtags into clusters and compared these with our manually defined categories.
- Measuring affinity and consistency: We evaluated to what extent ChatGPT's suggested hashtags aligned with actual co-occurrence data on Instagram, noting where the model favored semantic similarity over empirical relationships.

This multi-step methodology allowed us to assess the validity, consistency, and analytical productivity of synthetic data in social media research.

5. Findings

Our analysis revealed key differences between synthetic data generated by ChatGPT and real-world data collected from Instagram. The findings highlight the model's tendencies in determining relatedness between hashtags and its limitations in accurately reflecting real-world social media patterns.

1. Divergence Between Synthetic and Real-World Hashtag Occurrence

One of the most striking findings was that the majority of Instagram posts suggested by ChatGPT either did not exist or did not contain the related hashtags generated by the model. This suggests that ChatGPT does not base its responses on real-time social media data but rather on patterns inferred from its training data, which may not reflect actual usage on Instagram.

While ChatGPT can generate convincing hashtag lists, these lists often fail to align with actual co-occurrence patterns found in real Instagram posts. This raises questions about the reliability of synthetic data for studying social media trends and whether LLM-generated insights can serve as a proxy for empirical data collection.

2. Semantic Similarity vs. Empirical Co-Occurrence

ChatGPT's approach to determining related hashtags appears to rely more on semantic similarity rather than empirical co-occurrence within real social media posts.

This finding indicates that LLMs prioritize logical thematic associations over real-world hashtag behaviors, making them useful for understanding semantic relationships but less reliable for tracking actual social media trends.

3. Comparing Manual and Synthetic Cluster Labeling

To evaluate whether synthetic data could still be analytically useful, we compared the hashtag clusters generated by ChatGPT with those identified through manual coding. Our results showed that:

- 4 out of 6 thematic clusters generated by ChatGPT had strong affinities with manually coded clusters.
- While the categories overlapped significantly, ChatGPT's clusters often grouped hashtags based on conceptual relevance rather than frequency of co-occurrence in Instagram posts.

These results suggest that while synthetic data does not always match empirical data, it can still be productive for exploratory research, especially in developing initial categorizations and thematic groupings.

4. ChatGPT's Rationale for Related Hashtags

Another key finding was how ChatGPT explained the relationships between suggested hashtags. The model's reasoning was primarily based on broad thematic connections rather than real-world co-occurrence patterns.

This suggests that while ChatGPT can generate coherent justifications for relatedness, these justifications do not always reflect real-world social media interactions. Its explanations rely on logical thematic reasoning rather than empirical data-driven relationships.

5. Summary of Findings

- ChatGPT's suggested Instagram posts were often nonexistent or contained hashtags that did not match its own generated lists, indicating a gap between synthetic and real-world data.
- The model prioritizes semantic similarity over actual co-occurrence patterns, meaning it groups hashtags based on conceptual relevance rather than empirical social media usage.

- 4 out of 6 hashtag clusters created by ChatGPT aligned with manually identified clusters, suggesting some level of analytical usefulness.
- ChatGPT's reasoning for relatedness was coherent but not based on real-world data, often explaining connections in terms of broader themes rather than actual Instagram hashtag behaviors.

These findings demonstrate both the limitations and potential applications of synthetic data in social media analysis. While ChatGPT does not accurately replicate real-world co-occurrence patterns, it can still provide useful thematic insights and exploratory categorizations that may serve as a starting point for deeper empirical investigations.

6. Discussion

Our findings highlight both the potential and limitations of synthetic data in social media research. While ChatGPT-generated data can offer insightful thematic connections, it does not fully capture the complexity and nuances of real-world hashtag co-occurrence on Instagram. This divergence raises important questions about the reliability and applicability of synthetic data for studying online social phenomena.

1. Limitations of Synthetic Data in Representing Social Media Dynamics

One of the key limitations of synthetic data is its reliance on semantic similarity rather than empirical co-occurrence. ChatGPT generates related hashtags based on thematic relevance rather than real-world usage patterns, meaning that while its suggestions may appear logical, they often fail to reflect actual trends observed in Instagram posts. This can lead to misleading conclusions if researchers assume that synthetic relationships mirror real-world interactions.

This distinction is important for research that aims to understand how ideas and narratives spread on social media, as relying on synthetic data alone could obscure meaningful trends. Moreover, because LLMs do not have real-time access to social media platforms, they cannot track emerging trends, virality, or the shifting use of hashtags over time.

2. Value of Synthetic Data for Exploratory and Thematic Analysis

Despite its limitations, synthetic data can still serve as a valuable tool for exploratory research. Our study found that 4 out of 6 thematic clusters produced by ChatGPT overlapped significantly with manually identified clusters, suggesting that LLMs can provide a useful starting point for categorization and trend identification.

For researchers analyzing large amounts of social media data, synthetic hashtag generation could be used as:

- A preliminary step for identifying thematic groupings before conducting manual validation.
- A supplementary tool for hypothesis generation, helping researchers refine their questions before collecting empirical data.
- A way to explore conceptual relationships between topics, particularly in areas where empirical data may be sparse or difficult to access.

For instance, in the study of social imaginaries constructed through work-related memes, ChatGPT's generated hashtags could help map out different ideological positions (e.g., pro-work vs. anti-work sentiment) before conducting a deeper analysis of real-world posts.

3. Implications for AI-Assisted Research in Social Media Studies

Our findings contribute to the broader discussion on the role of AI in social media analysis. As AI-generated content becomes more prevalent, researchers must critically assess when and how synthetic data can be used effectively.

- Synthetic data is not a replacement for empirical research: While LLMs like ChatGPT can generate structured outputs quickly, they should be validated against real-world data before drawing conclusions.
- Understanding AI biases: Since LLMs do not have direct access to social media platforms, their outputs reflect patterns in their training data rather than real-time social dynamics. This means they may reinforce existing narratives rather than identify novel trends or unexpected co-occurrences.
- Potential for hybrid approaches: Combining AI-assisted analysis with traditional qualitative and quantitative methods could enhance research efficiency. For example, researchers could use LLMs for preliminary categorization, followed by automated data collection and manual validation to ensure accuracy.

7. Conclusion

Our findings suggest that synthetic data generated by LLMs like ChatGPT, while valuable for early-stage research and hypothesis generation, has limitations in accurately representing real-world social media dynamics. The primary limitation lies in its reliance on semantic similarity rather than actual co-occurrence data, which can result in a disconnect from real-world trends. However, the overlap between categories identified by ChatGPT and manual coding shows that LLMs can still be useful for thematic exploration. Future research should focus on improving LLM models by integrating real-time social media data, allowing them to better reflect actual trends and co-occurrence patterns. Enhancing AI with live data scraping or contextual understanding of hashtag usage could make synthetic data more reliable. Additionally, combining AI-generated insights with manual validation could improve accuracy and productivity in social media research. Hybrid approaches, where AI and human expertise work together, offer an efficient way to analyze large datasets while ensuring the quality of findings. Ultimately, future research should aim to bridge the gap between synthetic and real-world data to create more holistic and accurate methods for understanding social media behaviors.

8. References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

Birhane, A., Kalluri, P., Card, D., Hooker, S., & Barocas, S. (2021). "The Values Encoded in Machine Learning Research." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 173–184.

Hao, K. (2021). "What AI Still Can't Do." *MIT Technology Review*.

Jordon, J., Yoon, J., & van der Schaar, M. (2018). "Measuring the Quality of Synthetic Data for Use in Autonomy and Privacy." *arXiv preprint arXiv:1806.11345*.

Rubin, D. B. (1993). "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics*, 9(2), 461–468.