Propuesta de Trabajo de Fin de Grado

Modelos de lenguaje cuantizados para la generación eficiente de texto sintético

Laura Santamaría Báez
202105586@alu.comillas.edu

Tutor: Jenny Alexandra Cifuentes Quintero

Universidad Pontificia Comillas
E-3Analytics

1 Objetivos Generales y Específicos

1.1 Objetivo General

El presente Trabajo de Fin de Grado en Business Analytics tiene como objetivo principal desarrollar y proponer una metodología basada en la cuantización de modelos de lenguaje a gran escala (en adelante LLMs) y en el uso de técnicas de prompting con el fin de generar datos textuales sintéticos de manera eficiente, evaluando su calidad, diversidad y utilidad en tareas posteriores de procesamiento de lenguaje natural.

1.2 Objetivos Específicos

- Contextualizar la importancia de la generación de textos sintéticos en el ámbito del procesamiento de lenguaje natural, destacando su papel en la ampliación de *datasets*, la mitigación de desbalances de clases y la mejora de modelos supervisados.
- Revisar sistemáticamente las técnicas disponibles de cuantización de modelos y de *prompting* aplicadas a la generación de texto, identificando enfoques, ventajas y limitaciones reportadas en la literatura reciente.
- Diseñar un *pipeline* metodológico que integre la cuantización de LLMs y estrategias de *prompting* para la generación eficiente y controlada de datos textuales sintéticos.
- Evaluar los resultados del *pipeline* propuesto mediante métricas intrínsecas y extrínsecas, analizando tanto la calidad y diversidad del texto generado como su utilidad práctica en tareas de Procesamiento de Lenguaje Natural.

2 Motivación

El éxito y rendimiento de los LLM está intrínsecamente ligado a la disponibilidad de un gran volumen de datos diversos y de alta calidad para su entreno y evaluación. Sin embargo, la tasa de crecimiento de los datos de alta calidad es significativamente inferior al rápido aumento de los conjuntos de datos necesarios para su entrenamiento, situación que se agrava aún más con el exponencial aumento del tamaño de los modelos que cada vez necesitan más datos para asegurar una generalización robusta en distintas tareas y dominios. La obtención de dichos datos plantea importantes retos, debido a los elevados costes que su recopilación supone y a los problemas que plantean las preocupaciones en materia de privacidad o sesgos [Wang et al., 2024]. Si esta tendencia continúa, llegará un momento en el que no habrá suficientes datos de calidad para entrenar los modelos, lo que implica que, a no ser que se mejore la eficiencia de los datos o se descubran nuevas fuentes, el riesgo de *overfitting* por la escasez de datos de entrenamiento aumentará y el el crecimiento de los LLM podría ralentizarse considerablemente. En este contexto, los datos sintéticos surgen como una solución prometedora y esencial [Ding et al., 2024].

En esta línea de investigación, numerosos estudios se han enfocado en examinar el potencial de los LLM en la generación de datos sintéticos de entrenamiento para diversas tareas de procesamiento de lenguaje natural (NLP), principalmente mediante el uso de técnicas de *prompting*. En un inicio, se realizó una primera aproximación al uso de *prompting*, centrándose en una mejora en términos de diversidad, sesgo y eficiencia, comenzando con *prompts* más simples, como los *simple class-conditional prompts*, hasta otros más completos, como los *attributed prompts* que permitían incluir más de un atributo [Yu et al., 2023]. Más adelante, surge la necesidad de consolidar las distintas técnicas de *prompting*, concretamente en el contexto de los LLMs pre-entrenados y estandarizar lo que se

entiende como *datos de alta calidad* (en este caso, fidelidad y diversidad). En este punto, se empieza a generar datos mediante el *prompt engineering* y el *multi-step generation*, identificando la capacidad de los LLMs de seguir instrucciones [Long et al., 2024]. Actualmente, se han posicionado los LLM como *universal data augmenters*, por su gran capacidad de generar datos sintéticos. La investigación ahora se centra también en analizar la idoneidad de las diferentes técnicas de *prompting*, que van desde *zero-shot* y *few-shot* hasta técnicas más avanzadas como *topic-controlled prompting* o *iterative and feedback-driven generation* [Nadas et al., 2025]. Estos avances muestran una transición desde enfoques experimentales y fragmentados hacia marcos metodológicos más generalizables, lo que refuerza la relevancia de la generación de datos sintéticos como solución estratégica frente a la escasez de datos reales de alta calidad.

En este marco de avances metodológicos, el presente trabajo se orienta a contribuir a dicha línea de investigación mediante el diseño de una propuesta que combina la cuantización de LLMs con el uso de técnicas de *prompting* para la generación de datos textuales sintéticos. El objetivo es no solo explorar la viabilidad de integrar ambas estrategias como vía para reducir costes computacionales y mejorar la eficiencia en la generación, sino también evaluar de manera sistemática la calidad y la utilidad práctica de los datos producidos en tareas de procesamiento de lenguaje natural, ofreciendo así un enfoque que busca ser eficiente, reproducible y escalable.

3 Metodología Propuesta

La metodología de este trabajo de fin de grado se estructura en una serie de etapas diseñadas para responder a los objetivos planteados. En primer lugar, se realizará una revisión de la literatura centrada en el uso de LLMs para la generación automatizada de texto, con el propósito de identificar las técnicas de prompting más empleadas, los criterios habituales de evaluación (coherencia, relevancia y diversidad léxica) y las principales limitaciones reportadas en investigaciones previas. En una segunda etapa, se procederá a la selección de un modelo cuantizado ya entrenado, atendiendo a criterios de eficiencia computacional definidos por la relación entre el tamaño del modelo, los recursos necesarios para su ejecución y la precisión obtenida en las tareas de referencia. Posteriormente, se llevará a cabo la generación de datos textuales sintéticos en distintos formatos representativos de posibles aplicaciones (por ejemplo, descripciones de productos, resúmenes de opiniones de clientes o respuestas a correos electrónicos), empleando diversas estrategias de prompting que permitan controlar atributos como estilo, longitud o contenido temático. Una vez generado el corpus sintético, se desarrollará una evaluación de calidad basada en métricas automáticas que medirán coherencia, relevancia y diversidad, complementada con un análisis comparativo respecto a los estándares definidos en la literatura. Finalmente, los resultados obtenidos serán analizados y discutidos con el objetivo de contrastar los resultados con los estudios previos y extraer conclusiones sobre la utilidad práctica de los LLMs cuantizados en la generación de texto sintético.

4 Tabla de Contenidos propuesta

1. Introducción: (a) Objetivos (b) Motivación (c) Estructura del documento
2. Estado del arte
3. Metodología: (a) Selección de un LLM cuantizado (b) Prompting (c) Experimentación (d) Evaluación de la calidad del corpus sintético obtenido
4. Resultados
5. Conclusiones
6. Bibliografía

References

[Ding et al., 2024] Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

[Long et al., 2024] Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., and Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

[Nadas et al., 2025] Nadas, M., Diosan, L., and Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*.

[Wang et al., 2024] Wang, K., Zhu, J., Ren, M., Liu, Z., Li, S., Zhang, Z., Zhang, C., Wu, X., Zhan, Q., Liu, Q., et al. (2024). A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

[Yu et al., 2023] Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A. J., Krishna, R., Shen, J., and Zhang, C. (2023). Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784.