Exploration and selection of LLM models for financial text simplification

Exploración y selección de modelos LLM para la simplificación de texto financiero

Bertha C. Brenes-Brenes[1], Saul Calderón-Ramírez[2]

Brenes Brenes, B. C; Calderón-Ramírez, S. Exploration and selection of LLM models for financial text simplification. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 50-56.

https://doi.org/10.18845/tm.v37i7.7297

[1] Student. Instituto Tecnológico de Costa Rica, Costa Rica.
berthabb@estudiantec.cr
https://orcid.org/0009-0008-1303-7263

[2] Computer Scientist, Instituto Tecnológico de Costa Rica, Costa Rica.
sacalderon@itcr.ac.cr
https://orcid.org/0000-0001-9993-4388

Keywords

LLM models; simplification; SARI; BLEU; AI; Llama; financial data.

Abstract

This research is dedicated to the simplification of Spanish-language financial texts to enhance accessibility for screen readers. We present a qualitative and quantitative analysis of the text simplification process, employing a set of Spanish simplification rules and metrics. Our study evaluates the outcomes resulting from the application of three distinct financial datasets to four pre-trained models. The primary objective is to identify the most effective models for text simplification and determine those warranting further investment through fine-tuning and training. This study contributes to improving the accessibility and comprehensibility of financial documents for individuals with visual impairments.

Palabras clave

Modelos LLM; simplificación; SARI; BLEU; IA; datos financieros.

Resumen

Esta investigación está dedicada a la simplificación de textos financieros en español para mejorar la accesibilidad de los lectores de pantalla. Presentamos un análisis cualitativo y cuantitativo del proceso de simplificación de textos, empleando un conjunto de reglas y métricas de simplificación en español. Nuestro estudio evalúa los resultados obtenidos de la aplicación de tres conjuntos de datos financieros a cuatro modelos pre entrenados. El objetivo principal es identificar los modelos más eficaces para la simplificación de textos y determinar aquellos que justifican una mayor inversión para el fine-tunning y el entrenamiento. Este estudio contribuye a mejorar la accesibilidad y comprensibilidad de los documentos financieros para las personas con discapacidad visual.

Introduction

Large Language Models (LLMs) are powerful machine learning models that leverage extensive training data to understand, translate, generate,simplify or summarize text. These models offer versatility in their application and training methods. However, their usage typically demands significant computational resources due to their memory and processing requirements, often surpassing the capabilities of a single computer.

One prominent type of an LLM Modelos is the Transformer [1]. Transformer architecture excels at analyzing text to comprehend its context and generate human-like responses. It operates by breaking down input text into tokens and representing them as vectors, capturing both their meaning and context. This allows the users to interact with the model using natural language queries and instructions, harnessing the context for generating meaningful responses [2].

The central to the effectiveness of LLMs is the quality and relevance of the training dataset. In this study, we focus on a dataset consisting of financial documents, specifically tailored for individuals with visual impairments, already develop by another research team

Our primary objective is to simplify Spanish-language financial text to enhance the accessibility of paperwork in screen readers for visually impaired people. This is a priority in our project, because we stand with the equal accessibility of all the people. We can define simplification as the replacement of complex words to simpler words to make the sentences more simple and easy to understand. In this process it is needed to delete, add or replace some words. One short example of simplification is:

Complex text: *"In recent years, specifically in recent decades, communities have become a fundamental pillar of their own economy, of countries and even of the international economy."*

Simpler text: *"In recent decades, communities have become a pillar of the local and international economy."*

This paper aims to present a qualitative and quantitative analysis using different types of spanish simplification rules, and utilizing a range of simplification metrics to evaluate the results obtained from three distinct financial datasets and four pre-trained models. The goal is to identify the most effective models for simplifying text and determine which ones are worthy of further investment in terms of fine-tuning and training.

Theoretical framework

For this project we focus on the primary exploration of the models and their performance using metrics. Is important to explain the concepts of the models that we are using:

Llama Model: LLM model developed by Meta, we are using Llama 2-13bf. which is optimized for dialogue use cases. **Pegasus Model**: LLM Model developed by Google, is pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models.

FastChat Model: LLM Model developed by LMSYS, is an open platform for training, serving, and evaluating large language model based chatbots. **Alpaca Model**: LLM Model developed by Stanford University, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations.

Mectrics

After all the text has been applied to the models mentioned above it is necessary to evaluate the simplification. It was necessary to use metrics that work efficiently in Spanish and for a better understanding of their objective in the simplification the metrics will be presented divided on the category they work to evaluate the quality of each simplification.

First we have the metrics that evaluate the quality of text simplification:

- **BLEU**:This metric does not have a specific equation because it will depend on the n-gram which is actually a widely used concept from regular text processing and is not specific to NLP or Bleu Score. '"The BLEU metric is always a number between 0 and 1'[5]. This result value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts.
- **SARI**: The metric compares the predicted simplified sentences against the reference and the source sentences. Is defined by .The range of values goes from 0 to 100, the higher the value, the better the simplification is.

The second category is the lexical simplicity metrics for natural language. All these metrics work for Spanish text.

- **Fernandez huerta**: This metric is defined by , where μ is the average number of syllables per word and Fμ the average number of words per phrase. It goes from 0 to 100, where 100 is easier to read.

- **Szigriszt pazos**: This metric is defined by . It goes from 0 to 100 where 0 is the hardest. And contrary as in the Fernandez huerta, it explains how hard it is to read a sentence.
- **Gutierrez poloni**: proposed as a novel readability formula from scratch for Spanish, where: L is the total number of characters. A higher value indicates a higher readability.
- **Crawford**: Is defined by the formula , OP, the number of sentences per hundred words; SP, the number of syllables per hundred words, This formula returns an estimate of the years of schooling required to understand the text[3] The lower the value indicates an easy readability.

The third category evaluates simple to complex text comparison. In addition to the following metrics, various features like syllables, word count, polarity (polo), monosyllable count, exact copies, and additions/deletions proportion are used to evaluate.

- Compression ratio: Calculates the ratio of the number of characters between the simple and complex sentences, is defined as . The lower this metric, the more simple it is.
- Levenstein similarity: this metric measures the Levenshtein distance between two text segments as the minimum number of single character edits (insertions, deletions or substitutions). A lower value indicates more lexical transformations.
- Sentence split: Is defined as .It corresponds to the ratio between the number of sentences in the simple text segment and the number of sentences in the complex text segment , thus: The lower this metric, the more complex the sentences is
- Lexical complexity score: Is a task in NLP and computational linguistics that involves assessing the level of complexity of the vocabulary and word usage in a given text[4]. A higher value indicates a good simplification and is easier to understand.

Methodology

For this research project, we will use a dataset that has been developed for a research group. They extract information from 4 financial books and divide this text in segments, then evaluate manually the complexity of each segment in order to create a several dataset where the biggest count is more than 5000 complex texts, this dataset also contain a manual simplification and a gpt3 simplification.We can understand this better with the following example:

Book	Complex	Simple
15654_LibroBAC.pdf	Este se da cuando el asegurado llega donde el médico, recibe la atención, efectúa el pago directamente y después solicita a la aseguradora el reembolso.	Esto pasa cuando el asegurado recibe la atención médica, efectúa el pago y después solicita a la aseguradora el reembolso.

Therefore we structured our approach to this dataset into three distinct stages:

Stage 1: Model Exploration: In this initial phase, we began with an exploration of various language models. In the research group, there was an initial exploration with 13 complex texts using models such as GPT-3, Llama13, Alpaca 13b, FastChat, and mt5 in the web chat.lmsys. Our exploration aimed to delve deeper into these models by accessing their Hugging Face pre-trained versions in Spanish so we can explore their usage in python, how it responds to the prompt instruction and the computer resource usage.

a. **Model Selection:** We identified one or more models for each of the four mentioned. Our selection criteria included evaluating the quality of documentation, scientific or blog references, Python compatibility, and computational resource requirements.

b. **Text Testing:** We used 13 complex texts to assess the four selected models. This involved applying the BLEU and SARI metrics to generate a range of scores for each of the 13 texts.

c. **Manual Validation:** Additionally, we manually validated the simplifications based on established guidelines. Each text was rated on a scale from 1 to 5, with 5 representing the highest quality simplification.

Stage 2: Large-Scale Testing: In the second stage, we scaled up our analysis by applying a dataset of over 5,000 complex texts to each of the models. The following steps were undertaken:

a. **Metrics Application:** We applied all the metrics mentioned in the theoretical framework, including SARI and BLEU, to assess the quality of simplifications.

b. **Replication:** To ensure robust results, we conducted more than five replications for each text and model, allowing for a more detailed and specific evaluation.

Stage 3: Shorter Dataset Analysis: The final stage involved the application of a shorter dataset containing fewer than 3,000 complex texts. This dataset was curated for its quality and relevance but kept the same structure that was already defined

Results

For the first stage The models Alpaca, Llama, and FastChat demonstrated positive performance in simplifying Spanish text, and each had their moments to stand out in different contexts. The application of rules was more consistent and accurate compared to the Pegasus model. In table 1 we can see the values apply to the simplification and it shows how the FastChat models represent the best. I consider that pegasus could improve if we use a specialized dataset and better training, because its main failure today is the number of words with grammatical and incoherent errors, so perhaps it will improve in a next stage with fine-tuning.

Table 1. Results from the manual evaluation. The values range from 1 to 5, the higher is 5.

Model	Total
FastChat	4.75
Llama	4.08
Alpaca	3.75
Pegasus	2.75

For the second and third stage we have the average of the results for BLEU and SARI. According with the characteristics of the implementation of SARI and BLEU we verify the metrics with different references, as we can see on the following tables

Table 2. quality of text simplification metrics in SARI and BLEU respectively.

Num. refs	Alpaca	Llama	FastChat	Pegasus
2.0	21.40	21.50	24.73	30.35

Num. refs	Alpaca	Llama	FastChat	Pegasus
3.0	21.33	21.42	24.73	30.57
4.0	21.06	21.16	24.79	31.34
5.0	21.19	21.31	25.76	31.75

Num. refs	Alpaca	Llama	FastChat	Pegasus
2.0	55.54	59.32	71.47	18.07
3.0	56.18	59.60	71.47	16.73
4.0	55.54	58.79	77.63	16.76
5.0	56.01	59.25	78.18	16.56

Table 3. lexical simplicity metrics, the higher the values the better, except for the Szigriszt_pazos.

Metrics	Alpaca	Llama	FastChat	Pegasus
Fernandez huerta	77.13	79.15	76.22	99.12
Szigriszt pazos	73.95	75.94	72.78	95.80
Gutierrez poloni	35.77	37.51	36.47	46.37
Crawford	3.86	3.71	3.89	2.30

Table 4. simple to complex text comparison metrics. the higher the values the better, except for the compression ratio and Levenstein similarity.syllable, word and polly are just informative, addition and deletion.

Metrics	Alpaca	Llama	FastChat	Pegasus
Compression ratio	1.09	1.01	1.00	0.43
Levenstein similarity	0.96	1.00	1.00	0.55
Sentence split	1.00	1.31	1.00	1.03
Lexical complexity	9.65	9.65	9.73	9.41
Additions proportion	0.08	0.02	0.00	0.07
Deletions proportion	0.00	0.00	0.00	0.63
syllable count	48.12	46.21	50.11	15.17
word count	28.59	27.68	29.53	9.41
poly count	5.11	5.11	5.91	1.54
SBert	0.96	1.0	1.0	0.72

Discussion

The comprehensive analysis of our research findings, as presented in Table 2, underscores the strong performance of the FastChat model in simplifying Spanish text. Pegasus, while showing promise in SARI, exhibited a substantial deficit in BLEU scores. Nevertheless, it excelled in lexical simplification, which is an important finding. In Table 4, where we compared the transformation of complex to simple text, FastChat once again demonstrated its prowess, even without additions or deletions in the simplification process. Alpaca and Llama also delivered commendable results, positioning them not far behind FastChat. Taking all these values into consideration, we can conclude that FastChat has demonstrated a better performance, making it a prime candidate for the fine-tuning process. In contrast, Pegasus may warrant reconsideration. It’s noteworthy that these results were obtained within the context of an exploration approach without fine-tuning. This offers an optimistic outlook for the subsequent stages of our research, with the potential for even more impressive outcomes.

References

[1] P. Menon. (2023). Introduction to Large Language Models and the Transformer Architecture

[2] V.Chaudhary.(2023).Transformers and LLMs: The Next Frontier in AI, ur(https://www.linkedin.com/pulse/transformers-llms-next-frontier-ai-vijay-chaudhary/)

[3] Legible. Fórmula de Crawford, url(https://legible.es/blog/formula-de-crawford/)

[4] K.North,M.Zampieri, M.Shardlow.(2023). Lexical Complexity Prediction: An Overview, url(https://dl.acm.org/doi/10.1145/3557885)

[5] K. Doshi(2021). Foundations of NLP Explained — Bleu Score and WER Metrics, https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b *Repo: https://github.com/BerthaBrenes/Text-Simplification-with-LLM*