Sumarização de opinião multidocumento para o português: comparando um método baseado em grafo com um LLM

Gustavo Sampaio Lima, Davi Fagundes Ferreira da Silva, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos – SP – Brasil

gustavo.sampaio@usp.br, davi_fagundes@usp.br, taspardo@icmc.usp.br

Abstract. *In this paper, we explore a graph-based method for multidocument opinion summarization for Portuguese. The method, which consists of an updated version of the well-known Opinosis method (Ganesan et al., 2010), has its results compared to those produced by a large language model, Mistral, when performing the same task for a small corpus.*

Resumo. *Neste artigo, exploramos um método baseado em grafo para sumarização de opinião multidocumento para o português. O método, que consiste em uma versão atualizada do conhecido Opinosis (Ganesan et al., 2010), tem seus resultados comparados aos produzidos por um grande modelo de língua, o Mistral, ao realizar a mesma tarefa para um pequeno corpus.*

1. Introdução

Com a quantidade de informação atual, a tarefa de sumarização de textos tem enorme relevância. Em Processamento de Linguagem Natural (PLN), há uma grande diversidade de métodos de sumarização, incluindo desde abordagens extrativas clássicas até, mais recentemente, o uso de grandes modelos de língua (no inglês, *Large Language Models* - LLMs) para produção de *abstracts*, como apontam Souza et al. (2024).

Neste artigo, relatamos um experimento de sumarização de opinião multidocumento para o português. Atualizamos um método já clássico de sumarização abstrativa baseada em grafos, chamado Opinosis (Ganesan et al., 2010), e comparamos seus resultados com aqueles produzidos pelo Mistral, um LLM *open source* bem difundido atualmente, usando um corpus de referência para o português, o OpiSums-PT (López Condori et al., 2015). Nossa proposta é investigar o desempenho do Opinosis, mais barato que um LLM, identificando suas potencialidades e limitações.

A seguir, na Seção 2, apresentamos brevemente o método Opinosis. Na Seção 3, descrevemos nosso experimento, os dados utilizados e os resultados alcançados.

2. O método Opinosis

Nosso método adapta o Opinosis pela adoção do modelo gramatical *Universal Dependencies* (UD) (de Marneffe et al., 2021), usado por mais de 150 línguas. Para obter as etiquetas morfossintáticas da UD, utilizamos o modelo pré-treinado Porttinari (Pardo et al., 2021; Duran et al., 2023) no UDPipe 2 (Straka, 2018) e desenvolvemos um conjunto

de 12 regras, formuladas como expressões regulares sobre sequências de etiquetas. Essas regras são projetadas para identificar os trechos sintaticamente bem formados e mais informativos dos textos, como construções canônicas de sujeito-verbo-objeto e frases nominais complexas, que servirão como candidatos para a construção do grafo que dará origem ao sumário/resumo. A Figura 1 ilustra um grafo direcionado construído para as frases "Bateria legal, gostei dms.", "A bateria é bem boa, muito legal.", "A tela é top, bateria boa tbm." e "A bateria é muito boa, adorei.". Cada nó no grafo representa uma palavra única junto de sua etiqueta, e as arestas preservam a ordem original das palavras em cada sentença. O algoritmo de sumarização busca por caminhos que possuam alta pontuação de redundância (computada pelo número de ocorrências) e sejam gramaticalmente válidos. No exemplo da figura, a alta coocorrência de "a", "bateria" e "é" cria um caminho com maior peso, levando o Opinosis a gerar o resumo final que inclui o trecho "[...] a bateria é bem boa, [...]".

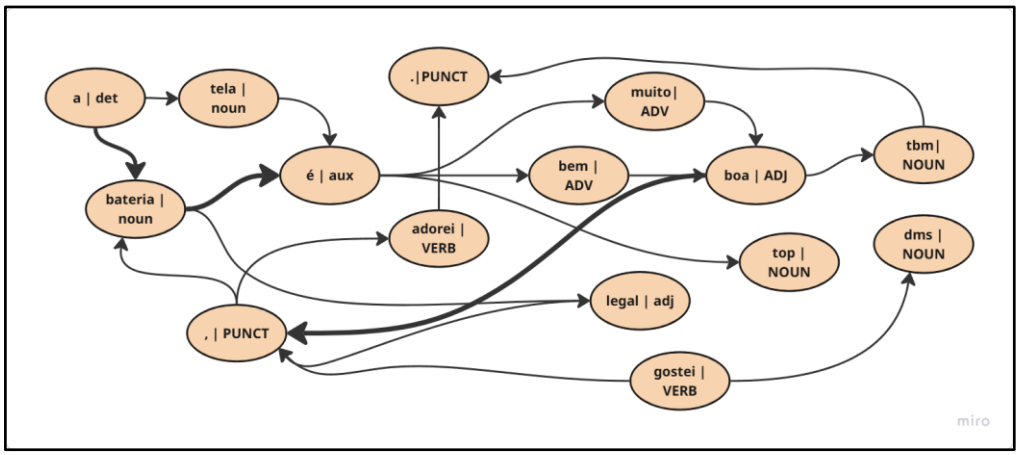

Figura 1. Exemplo de grafo produzido pelo método Opinosis adaptado

3. Experimento e resultados

Utilizamos o corpus OpiSums-PT, desenvolvido para a tarefa de sumarização de opiniões em português. O corpus inclui 17 tópicos, sendo 13 sobre livros e 4 sobre produtos eletrônicos. Para cada tópico, o corpus traz 10 opiniões de usuários e 10 resumos multidocumento criados por humanos, sendo 5 extrativos e 5 abstrativos. Neste trabalho, usamos esses resumos como referência, ou seja, "gabarito" (*gold standard*) em nossa avaliação, separando-os em função do tipo de resumo avaliado.

Nossa avaliação comparou o desempenho de dois sistemas distintos, o Opinosis adaptado e uma versão quantizada do modelo otimizado para instruções Mistral-7B-Instruct-v0.3 (Mistral AI Team; Jiang et al., 2023), para gerar resumos extrativos e abstrativos. Para as gerações com o LLM, utilizamos uma estratégia de decodificação gulosa a fim de garantir a consistência e a reprodutibilidade dos resultados. É válido ressaltar que, por conta de ser uma versão “menor” do modelo, essa redução pode impactar em sua performance. O modelo foi instruído a gerar um resumo de 2 a 4 sentenças, utilizando exclusivamente as 10 avaliações de cada tópico do corpus como fonte. A instrução (*prompt*) continha uma descrição clara da tarefa e exigia uma saída em formato JSON para facilitar o processamento automático.

A qualidade dos resumos gerados foi medida com a ROUGE (Lin, 2004) – em suas variantes ROUGE-1 e ROUGE-L – e com a BLANC (Vasilyev et al., 2020). A ROUGE é obtida pela contagem de n-gramas em comum entre o resumo automático e o(s) humano(s), enquanto a BLANC mede o quanto os resumos ajudam um modelo de língua a prever palavras ocultas em um texto, refletindo sua utilidade e coerência.

A Figura 2 mostra os resultados médios da ROUGE. O LLM se sai melhor na produção de resumos extrativos, mas perde para o Opinosis no caso dos resumos abstrativos. As hipóteses para explicar esse comportamento são que o LLM faz uma escolha melhor das sentenças que comporão o extrato e, no caso dos resumos abstrativos, o LLM é penalizado pela ROUGE (que privilegia sobreposição lexical com os resumos humanos de referência, situação em que o LLM tem desvantagem, pois este não necessariamente utiliza as mesmas palavras da referência humana).

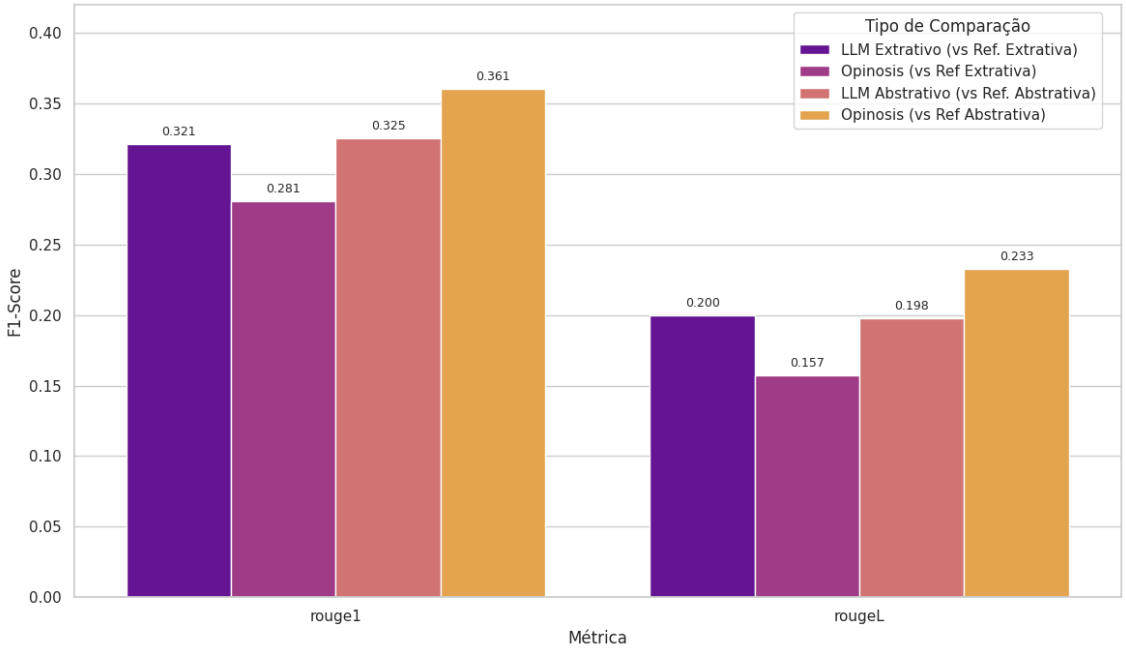

Figura 2. Análise de desempenho pela ROUGE (F1-Score)

A análise da métrica BLANC, exibida na Tabela 1, revela que o Opinosis obteve desempenho melhor, superando o LLM nas duas variações da métrica (*score* geral - *tune* - e alinhamento positivo - *help*).

Tabela 1. Resultados médios da métrica BLANC

Sistema	BLANC *Tune* (Média + DP)	BLANC *Help* (Média + DP)
Opinosis	**0,09 ± 0,08**	**0,13 ± 0,05**
LLM Extrativo	0,07 ± 0,07	0,12 ± 0,05
LLM Abstrativo	0,06 ± 0,07	0,11 ± 0,05

Este resultado indica que os resumos gerados pelo Opinosis são, em média, mais fiéis ao conteúdo original, provavelmente por sua abordagem ser ligada à extração e combinação de trechos originais, conseguindo manter uma conexão mais forte com o conteúdo fonte do que o LLM. Adicionalmente, o alto desvio padrão observado em todos os sistemas sugere que a qualidade final dos resumos foi fortemente influenciada pelas características das opiniões de cada tópico.

Como exemplo do material produzido automaticamente, mostramos resumos abstrativos gerados para o livro "Fala Sério, Amor!". Nota-se que o resumo do LLM é normalmente mais fluente, o que pode estar evidenciando uma das hipóteses anteriores, de que o LLM foi prejudicado na avaliação pelas métricas usadas. Se a avaliação abordasse qualidade linguística, ele potencialmente se sairia melhor.

- Opinosis: *ótimo livro pra quem quer dar boas risadas e com cada situação vivida por a protagonista; apaixona o público adolescente com as histórias de os namorados de nossa mais uma vez protagonista; tiver mais que 20 anos vai achar o livro, porém para adolescentes é realmente bom; bem engraçadas e com a linguagem descontraída e informal que thalita utiliza em suas obras.*
- Resumo abstrativo gerado por LLM: *O livro 'Fala Sério, Amor' é uma coleção de histórias divertidas e engraçadas sobre as aventuras amorosas de uma adolescente chamada Malu. As histórias são escritas em uma linguagem informal e descontraída, e retratam situações típicas que podem acontecer em uma vida adolescente, como se apaixonar por o melhor amigo, ficar com um loco em o carnaval, ir em a casa de o namorado por a primeira vez, entre outras. O livro é considerado engraçado e divertido, e é recomendado para adolescentes.*

Também foi conduzido um teste de concordância de polaridade de sentimentos (positiva ou negativa) entre resumos utilizando o TeenyTinyLlama-460m-IMBD (Correa et al., 2024), dada a importância desse tipo de informação ao se lidar com resumos de opiniões. Avaliaram-se separadamente os resumos abstrativos e extrativos gerados por LLM, comparando-os com a moda (polaridade mais frequente) dos resumos humanos correspondentes. Os resumos extrativos obtiveram 82,35% de concordância, enquanto os abstrativos alcançaram 88,24%, superando por ampla margem o Opinosis (70,59% e 58,82%, respectivamente).

Ao cruzar os resultados, emerge um claro *trade-off*. O LLM aparenta produzir resumos abstrativos com maior semelhança ao estilo humano e com mais concordância de polaridade, enquanto o Opinosis alcança resultados de avaliação melhores em termos das métricas ROUGE e BLANC. Assim, em certas situações, o Opinosis pode se mostrar uma alternativa interessante (por consistir em um método mais simples e computacionalmente mais barato).

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Também se agradece ao INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação) (processo CNPq/SECTICS/CAPES/FAPs #408490/2024-1) e à CAPES.

Referências

Correa, N.K.; Falk, S.; Fatimah, S.; Sen, A.; Oliveira, N. (2024). TeenyTinyLlama: open-source tiny language models trained in Brazilian Portuguese. Machine Learning With Applications, Vol. 16.

de Marneffe, M.C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021). Universal Dependencies. Computational Linguistics, Vol. 47, N. 2, pp. 255-308.

Duran, M.S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In the Proceedings of the 14th Symposium in Information and Human Language Technology (STIL), pp. 115-124.

Ganesan, K.; Zhai, C.; Han, J. (2010). Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In the Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 340-348.

Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L.R.; Lachaux, M.A.; Stock, P.; Scao, T.L.; Lavril, T.; Wang, T.; Lacroix, T.; Sayed, W.E. (2023). Mistral 7B. Disponível em https://arxiv.org/abs/2310.06825. Acesso em 23 de junho de 2025.

Lin, C.W. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In the Proceedings of the Text Summarization Branches Out Workshop, pp. 74-81.

López Condori, R.E.; Avanço, L.V.; Balage Filho, P.P.; Bokan Garan, A.Y.; Cardoso, P.C.F.; Dias, M.S.; Nóbrega, F.A.A.; Sobrevilla Cabezudo, M.A.; Souza, J.W.C.; Zacarias, A.C.I.; Seno, E.M.R.; Di Felippo, A.; Pardo, T.A.S. (2015). A Qualitative Analysis of a Corpus of Opinion Summaries based on Aspects. In the Proceedings of the 9th Linguistic Annotation Workshop (LAW), pp. 62-71.

Mistral AI Team (2024). Mistral 7B Instruct v0.3. [S. l.]: Hugging Face, 2024. 1 modelo de linguagem. Disponível em https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Acesso em 23 de junho de 2025.

Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pp. 1-10.

Souza, J.W.C.; Cardoso, P.C.F.; Paixão, C.A. (2024). Sumarização Automática. In H. M. Caseli e M. G. V. Nunes (eds), Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 3a edição, BPLN.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In the Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL), pp. 197-207.

Vasilyev, O.; Dharnidharka, V.; Bohannon J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. In the Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, p. 11-20.