Uso de modelos de linguagem na identificação de relações de coerência da *Rhetorical Structure Theory* em textos de *User-Generated Content*

Pedro Moreno Niella de Souza Sales Evangelista[1], Jackson Wilke da Cruz Souza[2]

[1]Instituto de Ciência, Tecnologia e Inovação - Universidade Federal da Bahia (UFBA)
Camaçari, Bahia / Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia
Salvador, Bahia / Brasil

{pedroniella, jackcruzsouza}@gmail.com

Abstract. *This work investigates the automatic identification of Rhetorical Structure Theory (RST) rhetorical relations in User-Generated Content, using Discourse Signals (DSs) as coherence evidence under the RST framework. The low coverage of classical connectives (19%) motivates a comparative pipeline with lexical heuristics, Support Vector Machine, and BERTimbau fine-tuning on the DANTEStocks corpus (4,048 financial tweets). BERTimbau achieved 95% accuracy and 93% weighted F1-score; however, class imbalance restricts the macro F1-score to 53%, showing that deep contextual modeling mitigates UGC linguistic noise but does not resolve the class imbalance problem.*

Resumo. *Este trabalho investiga a identificação automática de relações retóricas Rhetorical Structure Theory (RST) em User-Generated Content, utilizando Sinalizadores Discursivos (SDs) como evidências de coerência sob a ótica da RST. A baixa cobertura de conectivos clássicos (19%) motiva um pipeline comparativo com heurísticas léxicas, Support Vector Machine e fine-tuning do BERTimbau no corpus DANTEStocks (4.048 tweets financeiros). O BERTimbau alcançou 95% de acurácia e 93% de F1-Score ponderado; contudo, o desbalanceamento entre classes restringe o F1-Macro a 53%, evidenciando que a modelagem contextual profunda mitiga o ruído linguístico do UGC, mas não resolve o problema de desbalanceamento.*

1. Introdução

User-Generated Content (UGC) constitui uma das principais fontes de dados textuais de interesse para o Processamento de Linguagem Natural (PLN) [Gama e Souza, 2024]. Contudo, sua análise computacional esbarra em obstáculos como desvios linguísticos, variações ortográficas e fragmentação sintática, típicos de publicações em redes sociais [Di-Felippo et al., 2021]. No domínio financeiro, esses desafios se agravam pela brevidade das postagens e por tokens específicos, exigindo metodologias que superem a identificação de palavras-chave e alcancem a organização pragmático-discursiva das sentenças.

A *Rhetorical Structure Theory* (RST) [Mann e Thompson, 1988] oferece o suporte teórico necessário: a teoria decompõe o texto em *Elementary Discourse Units* (EDUs) e organiza as relações de coerência entre elas em estrutura hierárquica. Até o momento, a identificação dessas relações é majoritariamente feita com base em Marcadores Discursivos (MDs), especialmente, preposições e conjunções. Contudo, Das e Taboada (2019) demonstraram que a coerência textual (i) pode ocorrer a despeito de MDs comuns,

e (ii) sempre deixa evidências das relações no enunciado. Como resultado, os autores propuseram uma matriz de Sinalizadores Discursivos (SDs) que ampliam a noção de pistas de identificação das relações de coerência, considerando outros aspectos importantes para a sinalização dessas relações, como pontuação.

No cenário da língua portuguesa, esforços recentes têm mapeado a ocorrência desses sinalizadores em diferentes gêneros. Cardoso *et al.* (2024) e Gama e Souza (2024) sistematizaram a anotação RST para os domínios jornalístico e financeiro (este advindo de postagens de redes sociais), respectivamente, resultando em uma tipologia de SDs que indicam relações de coerência.

A automatização de algumas tarefas em PLN evoluiu de sistemas heurísticos rígidos para abordagens estatísticas e redes neurais. Conforme discutido por Monard e Baranauskas (2003), classificadores supervisionados tradicionais oferecem robustez ao operar sobre espaços vetoriais definidos por atributos extraídos de esquemas formais de anotação linguística [Leech, 1993]. O estado da arte foi impulsionado pelo surgimento de modelos baseados em *Transformers*, como o BERTimbau [Souza, Nogueira e Lotufo, 2020], capaz de capturar nuances sintáticas e semânticas contextuais.

Neste trabalho, propusemos e avaliamos um *pipeline* comparativo estruturado em três abordagens distintas: heurística, Aprendizagem de Máquina (AM) supervisionado clássico e a aplicação do modelo BERTimbau. O objetivo deste estudo é investigar em que medida a incorporação de SDs e de representações contextuais profundas permite classificar automaticamente relações RST em *tweets* financeiros.

Este estudo busca somar-se aos esforços históricos de análise discursiva automática para o português, como o *corpus* CSTNews [referência] e do *parser* discursivo DiZer [Pardo e Seno, 2005]. Como contribuição inovativa, apresentamos a expansão do escopo dos trabalhos para o domínio desafiador do UGC financeiro por meio de arquiteturas baseadas em *Transformers*.

2. Metodologia

O *pipeline* computacional para identificação e classificação de relações retóricas foi estruturado em ambiente Python, com as bibliotecas Scikit-Learn [Pedregosa *et al.,* 2011] e Hugging Face Transformers [Wolf *et al.,* 2020], em três abordagens: heurísticas léxicas, classificação supervisionada clássica e ajuste fino do modelo BERTimbau, aplicado ao *corpus* DANTEStocks. A variável-alvo compreende 12 relações RST mapeadas via rotulação semiautomática: cada sentença foi revisada em relação à tipologia de SDs proveniente do trabalho de Cardoso *et al.* (2024), e a relação foi atribuída conforme o conectivo detectado; sentenças sem SDs receberam o rótulo default *Elaboration*, seguindo a própria proposta de Mann e Thompson (1988). Essas 12 relações foram as únicas com ocorrências identificáveis via SDs explícitos no *corpus*. Embora os MDs tenham ocorrido em apenas cerca de 19% do *corpus*, a inclusão de sinalizadores gráficos e semânticos (como sinais de pontuação e menções de usuários) no *pipeline* heurístico elevou significativamente a cobertura, o que justifica a relação *Preparation* figurar como a classe majoritária, como demonstra-se na Tabela 1. Sentenças sem nenhum sinalizador detectado receberam o rótulo *default Elaboration*.

Na segunda abordagem, utilizamos o algoritmo *Support Vector Machine* (SVM) [Cortes e Vapnik, 1995] com *kernel* RBF para modelar uma matriz atributo-valor de baixa dimensão a partir dos *tweets*. Os 23 atributos foram organizados em quatro grupos conforme sua natureza linguística e funcional. Essa distinção é útil para mensurar, quais dimensões mais contribuem para a predição das relações RST. Os grupos são divididos

em *morfossintáticos absolutos* que são a contagem de tokens para 11 etiquetas advindas da *Universal Dependencies* [Nivre *et al.,* 2016]; *proporcionais, sendo valores relativos, calculados pela* razão entre a frequência de uma categorias críticas pelo tamanho da sentença; *morfológicos binários são* valores lógicos que indicam exclusivamente a presença ou ausência de flexões verbais (pretérito ou futuro) no campo *Feats*; e *estruturais/específicos cujo contêm* características próprias do domínio financeiro e da rede social, incluindo a posição relativa da sentença e a presença de hashtags, menções, URLs, tickers, percentuais e vocabulário da área. Para o AM, utilizamos validação cruzada de 5 pastas, com o hiperparâmetro `class_weight='balanced'` para lidar com o desbalanceamento de classes. Para o custo e a largura de banda do *kernel*, assumiu-se `C = 1.0` e `gamma = 'scale'`. Todos os outros hiperparâmetros foram mantidos *default* da biblioteca *scikit-learn*. A distribuição das relações RST neste estudo está da seguinte maneira: *Preparation* com 2.210 instâncias, *Elaboration* com 651, *Circumstance* com 451, *Condition* com 404, *Cause* com 85, *Contrast* com 84, *Concession* com 48, *List* com 41 e *Antithesis*, *Interpretation, Restatement, Background e Purpose com 74*

Como limite de desempenho, realizou-se o ajuste fino do modelo BERTimbau Base com 110 milhões de parâmetros. O tokenizador nativo processou os tweets com truncamento e *padding* limitados a 128 *tokens*. O dataset, então, foi particionado em 80% para treino (3.238 instâncias) e 20% para teste (810 instâncias). O treinamento ocorreu com GPU ao longo de 5 épocas, utilizando o otimizador AdamW com taxa de aprendizagem de $2x10^{-5}$. Além disso, foram definidos os seguintes hiperparâmetros para a calibração do modelo: `batch size` de 32, `weight decay` de 0.01, `warmup ratio` de 0.1 e `seed` igual a 42; para a avaliação final, baseamo-nos no conjunto de teste isolado.

3. Resultados

A avaliação contrastiva entre as abordagens evidencia o impacto da modelagem contextual em ambientes de UGC. Enquanto a abordagem heurística restringiu-se à presença explícita de SDs (cobertura de 19%), os modelos supervisionados estendem a classificação para a totalidade do *corpus*. O SVM alcançou acurácia total de 64% e F1-Macro de 31% na validação cruzada. O ajuste fino do BERTimbau estabeleceu o limite de desempenho do *pipeline*, atingindo 95% de acurácia total e 93% de *F1-Score* ponderado no conjunto de testes.

Com relação à análise de desempenho das classes de alta frequência, isto é, relações RST com maior número de instâncias no *corpus*, o BERTimbau concentrou-se nessas classes de maior suporte. A abundância de instâncias viabilizou a convergência do modelo, mas evidencia a necessidade de ampliar dados para relações subrepresentadas. A relação *Preparation* (442 instâncias) apresentou comportamento quase invariável, com Precisão de 98%, Revocação de 99% e *F1-Score* de 99%. Algo semelhante ocorreu com as relações *Circumstance* (*F1-Score* de 97% a partir de 90 instâncias), *Condition* (F1: 94%; 81 instâncias) e *Elaboration* (F1: 94%; 130 instâncias). Esse resultado reflete a regularidade dos SDs sintáticos associados a essas relações: conjunções como se e quando funcionam como âncoras de aprendizado, fornecendo ao BERTimbau pistas contextuais suficientes para distingui-las com alta confiança.

Quanto ao diagnóstico de erros nas classes de cauda longa, fenômeno em que poucas classes concentram a maior parte das instâncias enquanto muitas classes possuem frequência muito baixa [Buda, Maki e Mazurowski, 2018], o modelo apresentou

decréscimo ao lidar com relações retóricas de baixa frequência, resultando em um F1-Macro total a 53%. No teste, as relações *Antithesis* e *Interpretation* (ambas com apenas 4 instâncias cada) zeraram em todas as métricas de avaliação automática. Já as relações *Concession* e *List*, com 10 e 8 instâncias respectivamente, apresentaram *F1-Score* de 18% e 20%, sugerindo baixa eficácia para essas classes. Esse resultado é explicado, sobretudo, pela baixa quantidade de exemplos no *corpus*, já que relações RST com menos de dez ocorrências não fornecem variabilidade linguística suficiente para distingui-las com exatidão. Consequentemente, o classificador colapsou as predições raras em direção às classes majoritárias, a saber *Elaboration* e *Preparation*.

4. Considerações finais

Neste trabalho avaliamos a identificação automática de SDs no *corpus* DANTEStocks sob o escopo teórico da RST, contrastando heurísticas lexicais, SVM e o modelo BERTimbau. Os resultados dos experimentos sugerem que dicionários de conectivos são insuficientes para UGC. Além disso, o ajuste fino do BERTimbau demonstrou que arquiteturas de atenção conseguem generalizar além dos MDs explícitos, capturando padrões discursivos que escapam a abordagens léxicas. Ainda, o desbalanceamento quantitativo entre as relações RST restringiu o F1-Macro global, colapsando as categorias minoritárias.

Cabe destacar a circularidade metodológica inerente a este estudo preliminar. O fato de o BERTimbau ter sido treinado em um *corpus* com rotulação semiautomática (*silver-standard*) estritamente baseada nos SDs, a alta acurácia (95%), atesta-se a capacidade do modelo em aprender e replicar as regras heurísticas, e não necessariamente a concordância com a anotação que o humano faria (tida como *gold-standard*). Assim, a validação manual de uma amostra para aferir a real generalização do modelo será uma etapa essencial prevista para trabalhos futuros.

Ainda sobre trabalhos futuros, destacamos a necessidade de investigar métodos para a mitigação desse cenário de cauda longa, englobando técnicas de aumento de dados textuais (*data augmentation*), funções de perda sensíveis ao custo (*cost-sensitive learning*) e estratégias de *few-shot learning* via modelos de linguagem de grande escala. Ainda, pretendemos, de igual modo, aplicar o *pipeline* em outros *corpora* de UGC para validar a capacidade de generalização dos classificadores discursivos em múltiplos domínios informais da *Web.* Por fim, caberá avaliações qualitativas para mensurar linguisticamente a amplitude e profundidade dos resultados obtidos neste estudo preliminar.

Agradecimentos

Este projeto 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. Além disso, tivemos apoio de recursos financeiros da Chamada FAPESB/CNPq 004/2023 – Programa Primeiros Projetos.

Referências

Buda, M., Maki, A. and Mazurowski, M. A. (2018) "A systematic study of the class imbalance problem in convolutional neural networks", *Neural Networks*, v. 106, p. 249-259.

Cardoso, P., Souza, J., Rodrigues, R., Dantas, E., Bárbara, L. S., Araújo, M., Gama, N., Almeida, T., Cruz, G. (2024) "A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos", STIL 2024, p. 247–256.

Cortes, C. and Vapnik, V. (1995) "Support-vector networks", *Machine Learning*, v. 20, n. 3, p. 273-297.

Das, D. and Taboada, M. (2019) "Multiple signals of coherence relations", *Discours*, v. 24, p. 1-38.

Di-Felippo, A. *et al.* (2021) "Descrição preliminar do corpus DANTEStocks", In: *Anais do XIII Brazilian Symposium in Information and Human Language Technology (STIL 2021)*, p. 210-214.

Gama, N. S. e Souza, J. W. C. (2024) "Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuário", STIL 2024, p. 418–423.

Leech, G. (1993) "Corpus annotation schemes", *Literary and Linguistic Computing*, v. 8, n. 4, p. 275-281.

Mann, W. C. and Thompson, S. A. (1988) "Rhetorical Structure Theory: Toward a functional theory of text organization", *Text & Talk*, v. 8, n. 3, p. 243-281.

Monard, M. C. and Baranauskas, J. A. (2003) "Conceitos sobre aprendizado de máquina", In: *Sistemas Inteligentes: Aplicações a Modelos Biológicos e Biomédicos*, p. 89-114.

Nivre, J. *et al.* (2016) "Universal Dependencies v1: A Multilingual Treebank Collection", In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, p. 1659-1666.

Pedregosa, F. *et al.* (2011) "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, v. 12, p. 2825-2830.

Souza, F., Nogueira, R. and Lotufo, R. (2020) "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In: *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020)*, p. 405-417.

Wolf, T. *et al.* (2020) "Transformers: State-of-the-Art Natural Language Processing", In: Proceedings of EMNLP 2020: System Demonstrations, p. 38-45.