Anais do Congresso Brasileiro de Iniciação Científica
(https://periodicos.unisanta.br/COB)

ISSN 2317-1316

CIÊNCIA DE DADOS APLICADA AO PROCESSAMENTO DE LINGUAGEM NATURAL: UM ESTUDO DA APLICAÇÃO DE ÁLGEBRA DE CLIFFORD EM ANÁLISE DE SENTIMENTOS

APPLIED DATA SCIENCE IN NATURAL LANGUAGE PROCESSING: A STUDY ON THE APPLICATION OF CLIFFORD ALLGEBRA IN SENTIMENT ANALYSIS

André Leal Costa, Alexandre Garcia de Oliveira

FATEC Rubens Lara Unidade Baixada Santista
E-mail para contato: andre.costa01@aluno.cps.sp.gov.br

RESUMO – O Processamento de Linguagem Natural (PLN) enfrenta desafios recorrentes na tarefa de Análise de Sentimentos, principalmente em decorrência de ambiguidades semânticas, sutilezas contextuais e a ocorrência de ironias. Os métodos computacionais tradicionais frequentemente utilizam representações lineares ou abordagens procedurais que simplificam a estrutura geométrica subjacente à linguagem natural. Este trabalho propõe uma abordagem alternativa fundamentada na Álgebra de Clifford, modelando sentenças de textos como multivetores em um espaço com assinaturas geométricas bem definidas Cl(p, q, r). A metodologia consiste na extração de adjetivos a partir de postagens da rede social X por meio de processamento linguístico (NLTK), seguida pela polarização lexical (VADER) e vetorização via TF-IDF para geração dos coeficientes e dimensões do multivetor. Foram avaliadas quatro métricas de transformação do multivetor para o corpo dos Reais: a norma quadrada, o funcional linear geral, o traço matricial via representações de Pauli e a forma bilinear. Os resultados experimentais com um conjunto de 700 postagens indicam que todas as quatro abordagens apresentaram consistência na classificação de sentimentos, alcançando uma acurácia de 70%, o que demonstra o potencial da Álgebra de Clifford para tarefas de PLN. Por fim, são discutidas as limitações do polarizador lexical e propostas perspectivas futuras com bases de dados em língua portuguesa.

Palavras-chave: Álgebra de Clifford; Análise de Sentimentos; Processamento de Linguagem Natural; Classificação geométrica.

ABSTRACT – Natural Language Processing (NLP) faces recurrent challenges in Sentiment Analysis tasks, primarily due to semantic ambiguities, contextual subtleties, and the occurrence of irony. Traditional computational methods frequently employ linear representations or procedural approaches that simplify the geometric structure underlying natural language. This work proposes an

ISSN 2317-1316

alternative approach grounded in Clifford Algebra, modeling text sentences as multivectors in a space with well-defined geometric signatures Cl(p, q, r)$. The methodology consists of extracting adjectives from X (formerly Twitter) posts through linguistic processing (NLTK), followed by lexical polarization (VADER) and TF-IDF vectorization to generate the coefficients and dimensions of the multivector. Four transformation metrics from the multivector to the field of Real numbers were evaluated: the square norm, the general linear functional, the matrix trace via Pauli representations, and the bilinear form. Experimental results with a dataset of 700 posts indicate that all four approaches showed consistency in sentiment classification, achieving an accuracy of 70%, which demonstrates the potential of Clifford Algebra for NLP tasks. Finally, the limitations of the lexical polarizer are discussed, and future perspectives using Portuguese-language datasets are proposed.

Keywords: Clifford Algebra; Sentiment Analysis; Natural Language Processing; Geometry Classification.

1 INTRODUÇÃO

A Análise de Sentimentos é tradicionalmente descrita como um conjunto de procedimentos e modelos computacionais voltados para a detecção de emoções em dados textuais, visando categorizá-los sob polaridades específicas, como positivas, negativas ou neutras (Caseli e Nunes, 2024; Ramos e Freitas, 2019). Embora abordagens baseadas em aprendizado de máquina supervisionado e álgebra linear tradicional apresentem resultados funcionais, elas frequentemente privilegiam o aspecto puramente estatístico e procedural, mascarando as relações geométricas intrínsecas à estrutura semântica dos textos (Goodfellow et al., 2016). Esta questão se torna particularmente evidente em cenários de alta complexidade linguística, como em sentenças contendo ironia, nas quais classificadores comuns falham ao analisar termos isolados ("palavras brutas") em detrimento do contexto semântico global (Vieira e Lopes, 2010).

Como alternativa às limitações das abordagens lineares, a Álgebra de Clifford (também denominada álgebra geométrica) surge como uma ferramenta matemática capaz de estender os conceitos da álgebra linear clássica por meio de espaços com sinais bem definidos. A representação de uma sentença como um multivetor em um espaço Cl(p, q, r) permite que diferentes dimensões do texto capturem propriedades distintas: o espaço euclidiano p modela o sinal positivo, o espaço de Minkowski q modela o sinal negativo e o espaço isotrópico r atua na representação de elementos neutros (Silva, 1999).

ISSN 2317-1316

Dessa forma, a informação de erro ou de polarização emerge naturalmente da própria estrutura matemática do espaço geométrico. Inspirado na metodologia de análise estrutural aplicada a modelos compostos (Fong et al., 2019), este trabalho investiga como a modelagem em Álgebra de Clifford e suas diferentes métricas de projeção para números Reais podem viabilizar uma classificação robusta e matematicamente consistente para a Análise de Sentimentos.

2 MATERIAIS E MÉTODOS

A metodologia computacional e experimental adotada para este trabalho se fundamenta nas seguintes etapas estruturadas:

a) Aquisição e Limpeza de Dados: Foi selecionado um banco de dados composto por cerca de 700 postagens em língua inglesa estadunidense simulando publicações da rede social X. O pré-processamento envolveu a remoção de *emojis*, caracteres especiais e sinais de pontuação, além da conversão de todas as letras para minúsculo. Para maximizar a relevância semântica, utilizou-se o recurso *stopwords* da biblioteca *NLTK* para excluir as palavras gramaticais menos informativas.

b) Extração de Adjetivos: A identificação das classes gramaticais foi conduzida com o auxílio dos pacotes *averaged_perceptron_tagger_eng* e *punkt_tab* do *NLTK*, isolando os adjetivos contidos em cada publicação.

c) Polarização e Configuração de Assinaturas: Os adjetivos extraídos foram submetidos à análise do algoritmo *VADER*. As assinaturas do espaço Cl(p, q, r) foram definidas da seguinte forma: o parâmetro p recebeu a contagem de adjetivos positivos, o parâmetro q recebeu a contagem de adjetivos negativos, e o parâmetro r (espaço isotrópico) representou a quantidade de palavras significativas restantes da sentença, calculada pela relação total - (p + q).

d) Atribuição de Coeficientes: Para subsidiar a geração dos multivetores com valores numéricos robustos, utilizou-se a vetorização *Term Frequency-Inverse Document Frequency (TF-IDF)*. Os valores obtidos pelo *TF-IDF* para cada palavra correspondente foram passados como os coeficientes das dimensões do multivetor geométrico.

ISSN 2317-1316

3 RESULTADOS E DISCUSSÃO

O classificador fundamentado na Álgebra de Clifford foi submetido a testes experimentais com o corpus de 700 postagens pré-processadas, fornecendo as seguintes constatações:

a) Consistência Classificatória: Surpreendentemente, todas as quatro métricas de projeção para o corpo dos Reais apresentaram um desempenho idêntico de acurácia de 70%. Isso demonstra que, embora as métricas atribuam valores numéricos absolutos diferentes para expressar a intensidade dos sentimentos, a fronteira de classificação resultante da orientação geométrica da Álgebra de Clifford permanece altamente robusta e consistente.

b) Equivalência Teórica e Prática: Os resultados empíricos confirmaram as deduções matemáticas, indicando que o Funcional Linear Geral e o Traço Matricial produzem resultados numericamente idênticos, distinguindo-se da Norma Quadrada e da Forma Bilinear. No entanto, essas diferenças não introduziram anisotropias que pudessem comprometer a integridade semântica da classificação das postagens.

c) Análise de Limitações Lexicais: Observou-se que as poucas falhas de classificação não derivam do arcabouço geométrico em si, mas sim de limitações inerentes aos dicionários lexicais convencionais. A sentença "*a rainy day requires a cozy blanket and hot cocoa*" serve de ilustração: o polarizador *VADER* identificou "*rainy*" como fortemente negativo, superando o peso positivo de "*hot*" e rotulando incorretamente o sentimento geral da postagem. Este comportamento indica a necessidade de integrar representações semânticas contextuais em trabalhos futuros.

d) Comparação com o Estado da Arte: Semelhante ao trabalho de referência (Fong et al., 2019) no qual a estrutura categorial viabiliza composições flexíveis, a Álgebra de Clifford permitiu a introdução de diferentes representações métricas sem a necessidade de alterar a parametrização inicial do multivetor, provando sua versatilidade matemática.

4 CONCLUSÃO

Este trabalho demonstrou a viabilidade e a consistência matemática da aplicação da Álgebra de Clifford ao Processamento de Linguagem Natural (PLN), especificamente na

ISSN 2317-1316

modelagem geométrica para a Análise de Sentimentos em mídias sociais. Ao estruturar as sentenças textuais como multivetores em um espaço com assinaturas Cl(p, q, r), foi possível incorporar as dimensões positiva, negativa e isotrópica (neutra) diretamente na geometria da representação semântica, superando simplificações das abordagens lineares convencionais.

A avaliação experimental com um corpus de 700 postagens revelou que todas as quatro métricas de projeção testadas (Norma Quadrada, Funcional Linear Geral, Traço Matricial e Forma Bilinear) mantiveram uma acurácia consistente de 70%. Esse resultado confirma a hipótese de que a fronteira de classificação fundamentada na orientação do espaço geométrico de Clifford é robusta, independente da variação nos valores numéricos absolutos gerados por cada métrica de projeção. Além disso, a análise dos erros evidenciou que as principais divergências decorrem de limitações do dicionário lexical (*VADER*) na interpretação do contexto semântico global de termos sobrepostos, e não de falhas na formulação geométrica.

Como trabalhos futuros e desdobramentos desta pesquisa, propõe-se:

1. Integração de *Embeddings* Contextuais: Substituir ou complementar a extração baseada em dicionários lexicais estáticos e *TF-IDF* por modelos de *embeddings* dinâmicos baseados em arquiteturas de *transformers*, reduzindo ruídos em sentenças complexas ou irônicas.
2. Expansão para a Língua Portuguesa: Adaptar o pipeline de extração e polarização para corpora em português, investigando o comportamento das assinaturas do espaço Cl(p, q, r) frente às especificidades sintáticas e semânticas do idioma.
3. Exploração de Novos Espaços Geométricos: Avaliar o impacto do ajuste dinâmico de parâmetros nos subespaços euclidiano, de Minkowski e isotrópico para capturar com maior precisão nuanças e intensidades de sentimentos.

Em suma, a Álgebra de Clifford consolida-se como um arcabouço promissor e matematicamente rigoroso para o PLN, abrindo caminho para classificadores geométricos mais flexíveis e fundamentados.

5 DECLARAÇÃO DE USO DE IA

Declaração de Uso de IA: Durante a elaboração deste trabalho, utilizou-se a ferramenta Notebook LM para aprimoramento dos textos por mim elaborados. O conteúdo gerado foi re-

ISSN 2317-1316

visado, editado e validado pelos autores, que assumem total responsabilidade pelo texto final."

6 REFERÊNCIAS

CASELI, H. M.; NUNES, M. G. V. Processamento de linguagem natural: conceitos, técnicas e aplicações em português. 2. ed. [S.l.]: BPLN, 2024. pg. 10.

FONG, B.; SPIVAK, D. I.; TUYÊRAS, R. Backprop as functor: A compositional perspective on supervised learning. In: 2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS). IEEE, 2019. p. 1–13.

RAMOS, Barbara; FREITAS, Cláudia. Sentimento de quê?" uma lista de sentimentos para a Análise de Sentimentos. STIL, p. 15-18, 2019.

SILVA, Ana Paula da Cunda Corrêa da et al. Álgebras de Clifford: uma construção alternativa. Santa Catarina: Universidade Federal de Santa Catarina, 1999, pg. 27, 38.

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas em corpora. Rio Grande do Sul: EDIPUCRS, 2010. pg. 183.

WIKIPEDIA CONTRIBUTORS. Exterior algebra. Disponível em: <https://en.wikipedia.org/wiki/Exterior_algebra>. Acesso em: 30 de março de 2026.