UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ ROBSON DA SILVA ARAUJO JUNIOR

MINERAÇÃO DE POEMAS ATRAVÉS DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

CAMPINA GRANDE - PB

2021

JOSÉ ROBSON DA SILVA ARAUJO JUNIOR

MINERAÇÃO DE POEMAS ATRAVÉS DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Professora Dra. Lívia Maria Rodrigues Sampaio Campos

CAMPINA GRANDE - PB

2021

A663m Araujo Junior, José Robson da Silva.
Mineração de poemas através de técnicas de processamento de linguagem natural. / José Robson da silva Araujo Junior. – 2021.

11 f.

Orientadora: Profa. Dra. Lívia Maria Rodrigues Sampaio Campos.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Mineração de textos. 2. Processamento de linguagem natural. 3. Modelagem de tópicos. 4. Projeto Coletânea de poesias. 5. Part-of-Speech tagging. I. Campos, Lívia Maria Rodrigues Sampaio. II. Título.

CDU:004.439(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

JOSÉ ROBSON DA SILVA ARAUJO JUNIOR

MINERAÇÃO DE POEMAS ATRAVÉS DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professora Dra. Lívia Maria Rodrigues Sampaio Campos
Orientadora – UASC/CEEI/UFCG

Professor Dr. Elmar Uwe Kurt Melcher
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de maio de 2021.

CAMPINA GRANDE - PB

ABSTRACT

The contact with poems in basic education is an encouragement for students to discover the pleasure provided by the experience with poetic language. The *Coletânea de Poesias* project, held annually at *Fera Colégio e Curso*, is an initiative that aims to promote this contact through the reading, appreciation, and writing of poems, generating yearly a book with texts written by students of the secondary education. Given the manual analysis of these texts would be costly, the present work employed techniques of Natural Language Processing, such as Part-of-Speech tagging and topic modeling, to analyze the poems produced in the ten most recent editions of the project. The results obtained reinforce aspects related to the freedom of creation involved in poetic production and that the themes addressed by the students vary according to their maturity and their environment.

Keywords: Text mining; Natural Language Processing; Topic modeling; Poem; *Coletânea de Poesias*.

Mineração de poemas através de técnicas de Processamento de Linguagem Natural

José Robson da Silva Araujo Junior*
jose.robson.junior@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba

Orientadora: Lívia M. R. Sampaio Campos*
livia@computacao.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba

RESUMO

O contato com poemas na educação básica é um incentivo para que os alunos descubram o prazer proporcionado pela experiência com a linguagem poética. O Projeto *Coletânea de Poesias*, realizado anualmente no FERA Colégio e Curso, é uma iniciativa que se propõe a promover esse contato através da leitura, apreciação e escrita de poemas, gerando a cada ano um livro com textos redigidos por alunos dos ensinos fundamental e médio. Como a análise desses poemas seria custosa se feita manualmente, o presente trabalho empregou técnicas de Processamento de Linguagem Natural, como *Part-of-Speech tagging* e modelagem de tópicos, a fim de fazer a mineração dos textos produzidos nas dez edições mais recentes do projeto. Os resultados obtidos reforçam aspectos ligados à liberdade de criação envolvida na produção poética e que os temas abordados pelos alunos variam de acordo com a sua maturidade e o seu ambiente.

PALAVRAS-CHAVE

Mineração de textos. Processamento de Linguagem Natural. Modelagem de tópicos. Poema. *Coletânea de Poesias*.

1 INTRODUÇÃO

No contexto da educação básica, o trabalho com poemas na sala de aula é uma forma de estimular a sensibilidade e o gosto pela leitura nos alunos desde os anos iniciais de sua formação. A linguagem poética, em seus múltiplos significados, exercita no leitor a sua capacidade de interpretação, servindo não somente como instrumento para a apreciação desse gênero, como também para a interação do(a) estudante com o mundo [5]. Além da leitura, trabalhar com a escrita de poemas é uma forma de reforçar as características que compõem o gênero, levando à reflexão ativa sobre as escolhas que o perpassam [17]. Dessa forma, os alunos podem se tornar ainda melhores articuladores no uso da língua portuguesa.

Com o objetivo de promover o contato com o gênero poema através da leitura, apreciação e escrita, realiza-se, no FERA Colégio e Curso, na cidade de Patos - PB, o Projeto *Coletânea de Poesias*, que teve sua primeira edição em 2001. A cada ano, esse projeto tem como resultado a confecção de um livro contendo poemas escritos por alunos do 6º ano do ensino fundamental à 3ª série do ensino médio. Ao longo dessa trajetória, realizaram-se 19 edições, contemplando centenas de poemas de alunos que passaram pela escola.

Como os estudantes têm total liberdade quanto ao tema e à forma no momento da produção dos poemas para o projeto, registram-se vários assuntos que despertaram seus interesses nos textos publicados. Uma análise manual nesse contexto seria custosa, especialmente dada a extensão do corpo de texto já existente e pelo gênero muitas vezes exigir, em sua plurissignificação, uma leitura mais próxima para devida compreensão [20]. Ademais, dispor de uma análise automática também ajudaria a identificar outras tendências difíceis de serem avaliadas manualmente, como a forma dos textos e o vocabulário utilizado.

Para esses fins, pode-se recorrer ao Processamento de Linguagem Natural (PLN), uma área que fornece diversas ferramentas para a análise de linguagens naturais. Um exemplo de técnica de PLN é a modelagem de tópicos, que permite identificar agrupamentos de termos que costumam aparecer juntos, dando indícios de temas.

Dessa forma, o presente trabalho tem como objetivo realizar a mineração dos poemas produzidos no Projeto *Coletânea de Poesias*, com foco na identificação de temas comumente abordados pelos alunos, extração de características dos textos e avaliação de como esses aspectos se relacionam com as séries[1] de seus autores. A fim de atingir esses objetivos, elaboraram-se três perguntas para nortear a pesquisa: i) *Quais características textuais dos poemas ajudam a diferenciar a etapa de formação (série) dos alunos?*; ii) *Como os temas abordados pelos alunos mudaram de acordo com a etapa de formação (série)?* e iii) *Como os temas abordados pelos alunos mudaram de acordo com os anos (edições do projeto)?*

Como objeto de análise desta pesquisa, utilizou-se um recorte contemplando as dez edições mais recentes do projeto, publicadas entre 2010 e 2019. Além de servir como *corpus* para o trabalho, os textos presentes nessas edições também serviram para a exemplificação de conceitos na fundamentação teórica e na discussão.

Os resultados obtidos neste trabalho reforçam a questão da liberdade de criação envolvida na escrita poética, por não se perceberem padrões relacionados com as características dos textos escritos pelos alunos e suas séries. Além disso, percebe-se que certos fatores, como a maturidade e o ambiente, influenciam na variação de temas ao longo das edições do projeto e das séries dos alunos.

Inicialmente, na seção 2, discute-se a fundamentação teórica, apresentando-se definições utilizadas ao longo do trabalho; em seguida, a seção 3 lista alguns trabalhos relacionados e as suas contribuições para esta pesquisa; na seção 4, descreve-se o método empregado; a seção 5 traz os resultados e discussões; e, finalmente, na seção 6, discorre-se sobre as conclusões, limitações e trabalhos futuros.

[1] *Série*, nesta pesquisa, refere-se ao ano escolar do ensino fundamental ou à série do ensino médio.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção introduz conceitos essenciais sobre os quais o presente trabalho está pautado. Inicialmente, trata-se do gênero dos textos sob análise, poema, e dos seus aspectos que foram observados na extração de características textuais. Em seguida, comenta-se sobre a área de Processamento de Linguagem Natural, especificando-se as tarefas e conceitos que integram a metodologia desta pesquisa.

2.1 Poema

O poema[2] é um gênero textual que se caracteriza por explorar e expandir os limites da língua. A linguagem poética, portanto, diferencia-se da linguagem informativa e cotidiana por empregar palavras para compor imagens com significados que ultrapassam o concreto [4]. Por consequência, o texto poético estimula o leitor a explorar sua intuição, sua emoção e sua sensibilidade [26].

Um texto costuma ser reconhecido como poema por sua disposição gráfica, de forma que cada linha é denominada de *verso*. Apesar de essa não ser a única forma que um poema pode assumir, é o formato no qual a maioria dos textos do gênero se apresentam, incluindo os poemas sob análise. Agrupamentos de versos são chamados de *estrofes*, comumente associadas ao número de versos que comportam; *quadra* (ou *quarteto*), por exemplo, é a designação dada a uma estrofe formada por quatro versos. "Viagem poética" [16], apresentado abaixo, é um poema composto por três quadras:

> **Viagem poética**
>
> A poesia me leva
> Para uma terra distante
> Onde a felicidade
> Está em viver cada instante.
>
> Faz-me viajar o mundo
> Sem nem mesmo sair do lugar
> E ultrapassar fortalezas
> Com a força do sonhar.
>
> Leva-me com sutileza
> Pelas trilhas da emoção
> E sabiamente decifra
> Os enigmas do coração.

Ao se redigir um poema, é comum o emprego das *figuras de linguagem*, recursos estilísticos que ajudam a atribuir novos sentidos às palavras utilizadas, a exemplo da *metáfora*, uma comparação implícita entre dois conceitos. No caso de "Viagem poética", a própria noção de "viagem", mencionada no título, pode ser interpretada como uma metáfora, ao ser comparada com a experiência proporcionada pela poesia.

Os temas mais diversos podem motivar a escrita de um poema. Retomando "Viagem poética", tem-se um exemplo de metalinguagem, pois se utiliza o próprio poema para falar do efeito que a poesia causa em seus leitores. Temas podem ser extraídos até mesmo da trivialidade do cotidiano: em "Toda família tem" [27], como o título sugere, sua autora descreve pessoas que costumam integrar os núcleos familiares, como ilustrado em sua primeira estrofe ("Toda família tem / Uma avó boa de fogão / Um primo esquisito / E um tio "bebarrão.").

[2] Apesar de *poema* e *poesia* serem muitas vezes tratados como sinônimos, o *poema* se refere a um gênero textual, comumente disposto em versos; já a *poesia* se refere à subjetividade emocional manifestada no poema, mas não limitada a esse gênero.

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa que envolve o uso de diversas técnicas computacionais para análise e representação de linguagens naturais (como a linguagem falada), a fim de realizar diferentes tarefas, como a desambiguação de sentido de palavras, a síntese de fala e a análise de sentimentos [19]. Nesta seção, detalham-se os processos e tarefas comuns em PLN que foram utilizados neste trabalho: o pré-processamento do texto, a *Part-of-Speech tagging* e a modelagem de tópicos.

2.2.1 Pré-processamento de texto. Pré-processar um texto consiste em prepará-lo para análise, normalizando-o segundo o padrão com o qual se espera trabalhar. É uma etapa crucial para as demais tarefas em PLN, impactando diretamente a qualidade dos resultados encontrados [20]. O pré-processamento de texto pode envolver diversas etapas, dentre as quais se destacam a tokenização, padronização de capitalização (*case folding*) e remoção de *stopwords*.

A tokenização é o processo de transformação de um texto em unidades menores, denominadas *tokens*, tipicamente as palavras do texto. Dessa forma, os espaços em branco e outros delimitadores, como sinais de pontuação, são considerados os elementos separadores dos termos [13]. A manutenção ou não dos sinais de pontuação depende da tarefa de PLN que se deseja desempenhar, pois estes são importantes na delimitação das sentenças do texto.

A padronização da capitalização do texto consiste em convertê-lo todo para uma mesma caixa, alta ou baixa. Dessa forma, palavras grafadas como "AMOR", "Amor" ou "amor" são normalizadas à mesma representação ("amor", por exemplo), evitando tratá-las como se fossem palavras distintas. Assim como na tokenização, a transformação do texto todo para uma mesma caixa nem sempre é aplicada. Em alguns casos, por exemplo, pode-se desejar diferenciar nomes próprios de nomes comuns [13].

Certas palavras, quando isoladas de seus contextos, não exprimem significados determinantes na compreensão do texto. Em geral, é o caso de artigos como "a" e "o", por exemplo. Para algumas atividades de PLN, portanto, é comum selecionar um conjunto de palavras (chamadas de *stopwords*) e removê-las dos documentos, por não contribuírem significativamente para os resultados desejados [13]. A descoberta de *stopwords* pode ser iniciada a partir da avaliação das palavras que mais costumam ocorrer nos documentos, mas também há diversas listas de palavras comumente descartadas para cada idioma.

Além da alta frequência, a determinação de que palavras devem ser consideradas *stopwords* está atrelada também ao contexto de análise. Em certos contextos, palavras que costumam ser ignoradas podem ser unidades importantes de significado. Dessa forma, a utilização (ou não) de listas de palavras a serem desconsideradas deve ser bem avaliada [25].

Na Figura 1, utiliza-se o texto do poema "Identidade nacional" [7] para exemplificar um possível resultado da aplicação das etapas de pré-processamento supracitadas.

2.2.2 Part-of-Speech tagging. A tarefa de *Part-of-Speech tagging* (*POS tagging*) consiste em desempenhar uma análise morfológica

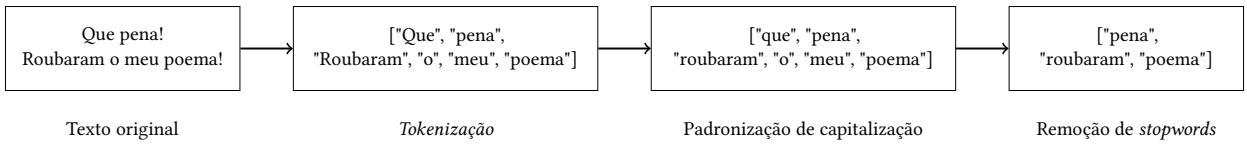

Figura 1: Etapas de pré-processamento no poema "Identidade nacional".

dos termos de um documento, rotulando cada um em função de sua classe gramatical [13]. Uma mesma palavra pode integrar diferentes classes, então um modelo que desempenha essa tarefa (conhecido como *POS tagger*) deve considerar tanto a sua definição quanto o contexto no qual ela está sendo empregada. Em geral, *POS taggers* são modelos estatísticos de Aprendizagem de Máquina e possuem uma alta acurácia na determinação das classes gramaticais [25]. A Figura 2 ilustra o resultado da aplicação dessa tarefa em um verso do poema "Identidade nacional".

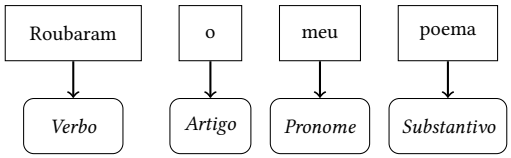

Figura 2: Exemplo da aplicação de *PoS tagging*.

2.2.3 Modelagem de tópicos. A modelagem de tópicos é uma tarefa comum em Aprendizagem de Máquina e Processamento de Linguagem Natural, tendo como objetivo descobrir grupos de termos que costumam aparecer juntos, comumente denominados *tópicos*. Trata-se de uma abordagem não supervisionada, ou seja, que tenta fazer a descoberta de padrões sem depender de os dados terem rótulos pré-definidos. Os resultados dessa tarefa são afetados por diversas decisões, incluindo os parâmetros passados (como a quantidade de tópicos esperados) e quais termos são filtrados anteriormente [3].

3 TRABALHOS RELACIONADOS

No gênero poema, Navarro-Colorado [20] utilizou modelagem de tópicos a fim de descobrir temas comuns nos sonetos da era de ouro espanhola, comparando as descobertas com os estudos feitos sobre os textos anteriormente e com a leitura dos poemas. Nesse trabalho, descobriu-se que alguns tópicos não representavam temas concretos, mas sim ideias que costumavam ser retratadas pelos autores ou grupos de palavras com características em comum. Ainda, discutiram-se formas de comparar os modelos automaticamente, testando diferentes técnicas para avaliá-los quantitativamente.

Junior, Rossi e Lobato [12] aplicaram modelagem de tópicos em letras de música para a descoberta de temas mais contemplados por cantores do sertanejo. Além disso, algumas características derivadas a partir dos textos foram utilizadas para comparar as letras de artistas masculinos e femininos, obtendo resultados que indicam a possibilidade de utilizar essas informações na diferenciação entre os gêneros.

Nunes, Souza e Cotait [21] aplicaram um algoritmo de agrupamento de textos curtos, *Gibbs Sampling for the Dirichlet Multinomial Mixture* (GSDMM), em *tweets* na língua portuguesa, conseguindo determinar diferentes assuntos em um conjunto de postagens relacionadas com desastres naturais. Similarmente, Dimitriadis [9] empregou GSDMM em *tweets* em grego escritos no contexto da pandemia da Covid-19, encontrando uma performance melhor, se comparada com outros algoritmos de modelagem de tópicos.

As descobertas descritas nos trabalhos supracitados dão indícios da aplicabilidade do Processamento de Linguagem Natural em diversos contextos, incluindo o trabalho com o gênero poema. A partir de suas constatações e das técnicas utilizadas para as suas respectivas análises, pautou-se o método aplicado na presente pesquisa, descrito na seção seguinte.

4 METODOLOGIA

Esta seção descreve a metodologia utilizada, detalhando os procedimentos que foram empregados para as finalidades deste trabalho. A pesquisa se deu em uma abordagem quantitativa, com objetivo descritivo e procedimento técnico experimental. O código, desenvolvido em Python[3], está disponível em um repositório aberto[4] na plataforma de hospedagem de código GitHub.

4.1 Recuperação dos dados

Apesar de todas as edições do *Coletânea de Poesias* estarem disponíveis no formato de livro físico, a equipe responsável pelo projeto cedeu os textos em versão digital no formato DOCX, especificamente para os propósitos da presente pesquisa. Esses arquivos seguem uma estrutura bem definida, o que facilitou o emprego da biblioteca python-docx[5] em sua leitura e extração de conteúdos. A partir desse ponto, utilizou-se a biblioteca pandas[6] para dispor os poemas em uma tabela contendo as suas informações principais: seu *título*, seu conteúdo (*texto*), o ano da *edição* que o contém, o nome do(a) *estudante* que o escreveu e a sua correspondente *série*. Foram totalizados 684 poemas, seguindo a distribuição por edição e série detalhada na Tabela 1.

4.2 Pré-processamento dos dados

Neste trabalho, os textos dos poemas foram utilizados na íntegra para as etapas de extração de características textuais e *PoS tagging*, uma vez que todas as palavras precisavam ser consideradas em ambas as atividades. Nesses casos, a etapa de pré-processamento consistiu apenas em operações mais simples, como a remoção de espaços excedentes. Como os textos utilizados eram provenientes de arquivos relativos aos livros publicados, eles já estavam devidamente padronizados.

[3]https://www.python.org/
[4]https://github.com/JRobsonJr/pln-poemas
[5]https://python-docx.readthedocs.io/en/latest/
[6]https://pandas.pydata.org/

Tabela 1: Distribuição de frequências de poemas por série e por edição

Edição \ Série	6°	7°	8°	9°	1ª	2ª	3ª	Total
2010	10	11	10	11	11	10	10	73
2011	6	12	13	10	10	10	10	71
2012	7	11	8	10	10	10	10	66
2013	11	13	9	12	11	10	10	76
2014	6	12	10	10	10	10	10	68
2015	11	8	9	11	10	10	10	69
2016	6	7	11	11	10	10	10	65
2017	7	7	6	10	10	10	10	60
2018	10	10	9	9	10	10	10	68
2019	9	8	11	10	10	10	10	68
Total	83	99	96	104	102	100	100	**684**

No caso da tarefa de modelagem de tópicos, foi necessário efetuar um pré-processamento com mais fases. Inicialmente, os textos foram convertidos para caixa baixa e tiveram seus caracteres não alfabéticos removidos. Em seguida, gerou-se uma lista de *stopwords*, com intuito de manter apenas os termos mais relevantes para a etapa de modelagem de tópicos. Esse conjunto de palavras foi concebido iterativamente, considerando o contexto da análise e observando as palavras muito frequentes. Em experimentos iniciais, esses termos costumavam aparecer na maioria dos tópicos e, por vezes, atrapalhavam na interpretação dos tópicos e/ou na sua devida separação.

Finalmente, a lista de *stopwords* utilizada foi composta por: verbos que são comumente utilizados como auxiliares ("ser", "ter", "haver", "estar" e "ir"), ou seja, verbos que não são o núcleo de sentido quando acompanham outros verbos (como em "está estudando" ou "vou pensar"); conectivos (conjunções e preposições); termos que substituem ou acompanham os substantivos (artigos e pronomes); e alguns advérbios muito frequentes, como "não", "assim" e "quanto".

4.3 Extração de características textuais

Cada um dos poemas foi descrito através de um conjunto de características textuais, sendo que algumas das quais são gerais, como o número de palavras e a diversidade lexical [12], e outras estão mais atreladas ao gênero poema, como o número de versos. As variáveis observadas e suas respectivas descrições estão detalhadas a seguir:

- *Número de palavras:* contagem de palavras do poema;
- *Número de palavras únicas:* contagem de palavras diferentes entre si no contexto do poema;
- *Diversidade lexical:* razão do número de palavras únicas pelo número de palavras;
- *Número de palavras raras:* contagem de palavras do poema que não ocorrem em outros poemas;
- *Taxa de palavras raras:* razão do número de palavras raras pelo número de palavras do poema;
- *Número de caracteres:* contagem de caracteres;
- *Média de tamanho de palavra:* razão do número de caracteres pelo número de palavras;
- *Número de versos:* contagem de versos do poema;
- *Média de palavras por verso:* razão do número de palavras pelo número de versos do poema;
- *Número de estrofes:* contagem de estrofes do poema.

4.4 *Part-of-Speech tagging*

Para a tarefa de *PoS tagging*, utilizou-se a biblioteca nlpnet[7], que apresenta uma acurácia de pelo menos 93,66% para a língua portuguesa [10]. Como os resultados apontados pela biblioteca são mais detalhados do que o necessário para a análise, eles foram mapeados nas classes gramaticais da língua portuguesa (substantivo, verbo, adjetivo, pronome, artigo, numeral, preposição, conjunção, interjeição e advérbio) e palavra denotativa. Para poder comparar os textos entre si, a contagem de cada classe foi dividida pelo total de palavras, representando a frequência relativa de cada uma.

4.5 Análise descritiva dos dados

A partir dos resultados do *PoS tagging* e das características textuais observadas, investigaram-se as suas distribuições de valores, buscando entender melhor o perfil dos poemas. Em seguida, visando investigar a possibilidade de associação entre esses valores e a série dos estudantes, executou-se o teste τ_b de Kendall para cada um deles. Trata-se de um teste de hipótese não paramétrico, em que a estatística de teste que ajuda a determinar a associação ou não entre duas variáveis é o coeficiente τ_b de Kendall. Esse teste foi considerado adequado para esse contexto, uma vez que se trata da relação entre uma variável ordinal (a série) e outra variável ordinal ou contínua (a característica observada) [14]. Utilizou-se, para tal, a implementação disponível na biblioteca scipy[8].

Nesse caso, considera-se que a hipótese nula H_0 é "a característica não está associada à série do aluno", enquanto a hipótese alternativa H_1 considera que "a característica está associada à série do aluno". Ao se aplicar o teste, obtém-se um coeficiente chamado τ_b de Kendall e um p-valor. A interpretação desses resultados segue da seguinte forma: caso o p-valor resultante seja menor ou igual ao nível de significância preestabelecido (adotou-se o valor padrão de 5%, ou seja, testa-se se p-valor <= 0,05), rejeita-se a hipótese nula. Assim, assume-se que o valor da associação é o coeficiente τ_b. Caso contrário (se p-valor > 0,05), aceita-se a hipótese nula, ou seja, assume-se que não há associação entre as variáveis.

4.6 Modelagem de tópicos

Como os poemas sob análise são, em geral, textos com poucas palavras, optou-se pela adoção do algoritmo *Gibbs Sampling for the Dirichlet Multinomial Mixture* (GSDMM) [28]. Esse algoritmo gera um modelo probabilístico que se mostra adequado no contexto de clusterização de textos curtos, servindo também para a modelagem de tópicos. Nesse caso, considera-se que cada documento está associado a um único tópico.

O GSDMM, ao realizar o agrupamento dos documentos, baseia-se principalmente nos valores de dois parâmetros. O primeiro, α, estabelece a influência do tamanho dos grupos, ajudando na manutenção de documentos que contêm as mesmas palavras em um mesmo grupo (princípio da completude). Por sua vez, o parâmetro β quantifica a influência das palavras utilizadas, visando manter documentos com palavras distintas em grupos distintos (princípio da homogeneidade). Ambos, α e β, assumem valores entre 0 e 1; no contexto desta pesquisa, testaram-se os valores arbitrários 0,05; 0,1; 0,25; e 0,5, gerando diferentes modelos para cada combinação.

[7] http://nilc.icmc.usp.br/nlpnet/
[8] https://docs.scipy.org/

O GSDMM tem como característica inferir automaticamente o número ideal de grupos a partir dos parâmetros fornecidos, exigindo apenas um valor máximo esperado K. Nesta pesquisa, escolheu-se fixar esse valor em 50, ao se perceber que os resultados naturalmente convergiam a números menores que esse. Por fim, há um parâmetro I que dita o número de iterações a serem executadas, ou seja, o número de vezes que os grupos são reorganizados. Para esse fator, escolheu-se o valor de 30, pois o algoritmo tem a característica de convergir com um pequeno valor.

Uma outra particularidade do GSDMM é que não se considera a repetição de palavras em um determinado documento, partindo da ideia de que textos mais curtos tendem a ser menos repetitivos. Portanto, além das etapas de processamento mencionadas anteriormente, removeram-se ocorrências adicionais de palavras em cada um dos textos.

4.6.1 Avaliação automática dos modelos. Como foram geradas diversas configurações de valores para α e β, seria inviável interpretar manualmente cada modelo produzido. Fez-se necessário, portanto, definir uma métrica para avaliá-los automaticamente. Para isso, recorreu-se à métrica *Normalized Pointwise Mutual Information* (NPMI), já proposta e aplicada anteriormente para a avaliação de modelos de tópicos [15].

NPMI é uma medida que quantifica o quão associadas são duas palavras, retornando um valor entre -1 (as palavras nunca ocorrem simultaneamente) e 1 (as palavras sempre ocorrem simultaneamente). A partir desse conceito, definiu-se que o *score* de um tópico é dado pela média dos valores de NPMI para cada par de palavras pertencentes ao tópico, considerando, nesse caso, as suas dez palavras mais frequentes. A intuição para esse cálculo está na ideia de que, quanto mais as palavras de um tópico estão associadas entre si, mais coerente o tópico parece ser. Finalmente, definiu-se um *score* para um modelo a partir da média ponderada dos *scores* dos tópicos que o compõem, com o peso de cada tópico correspondendo à quantidade de documentos que este contém. Assim, sugere-se que a coerência do modelo está associada à coerência de seus tópicos, com a média ponderada servindo para compensar os potenciais desbalanceamentos em seus tamanhos.

Dessa forma, utilizou-se o *score* definido para um modelo na seleção do melhor par de parâmetros de forma automática. É importante ressaltar, contudo, que os resultados encontrados não indicam necessariamente a qualidade dos modelos, servindo apenas como guia da escolha. Visando facilitar a compreensão do cálculo em maiores detalhes, produziu-se um material adicional[9], disponível no repositório do projeto.

4.6.2 Interpretação dos tópicos. Observar a coerência dos resultados gerados com a modelagem de tópicos é um desafio, pois, mesmo havendo métricas que ajudem nessa tarefa, não é possível assegurar a interpretabilidade dos resultados [15]. Apesar de haver uma métrica guiando a escolha, esse resultado não quer dizer que os tópicos são necessariamente coerentes e, além disso, a rotulação dos tópicos ainda deve ser feita manualmente.

Dessa forma, analisaram-se as dez palavras mais comuns em cada tópico, tentando determinar uma ideia que os resumisse, sendo também selecionados alguns poemas de cada um deles para avaliar essa escolha. Por fim, a partir dos tópicos descobertos, foi analisada a sua relação com a série e a edição dos poemas, promovendo-se uma discussão a respeito desses resultados.

[9] https://github.com/JRobsonJr/pln-poemas/blob/main/materiais/score-npmi.pdf

5 RESULTADOS E DISCUSSÃO

Nesta seção, apresentam-se os resultados dos passos detalhados na metodologia e a discussão de suas implicações em relação às perguntas de pesquisa elaboradas.

5.1 Características dos poemas

Cada poema sob análise tem uma média de aproximadamente 67,05 palavras, sendo, dessas, 45,69 únicas. A maior parte dos textos (cerca de 86,99%) contabiliza menos de 100 palavras, enquanto apenas 19 poemas têm mais de 150 palavras. A Figura 3 mostra a distribuição do número de palavras por poema, com uma linha vertical na mediana.

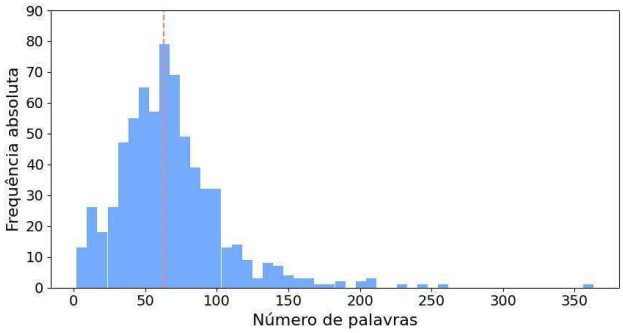

Figura 3: Distribuição do número de palavras por poema.

Considerando os aspectos específicos do gênero textual, cada poema tem, em média, 14,32 versos, distribuídos em 3,49 estrofes. As disposições de versos por estrofes (aqui referenciadas como "estruturas") mais comuns envolvem quadras, estrofes com quatro versos. A Tabela 2 mostra as cinco estruturas mais comuns nos poemas sob análise; 4-4-4, por exemplo, corresponde a um poema composto por três quadras. Embora haja poemas com as mais variadas estruturas (307 diferentes entre si), essas cinco concentram 25% dos poemas.

Tabela 2: Estruturas mais comuns de poemas

Estrutura	Frequência
4-4-4	53
4-4-4-4	42
4-4	33
4-4-4-4-4	25
4	18

A Tabela 3 apresenta as porcentagens de classes gramaticais nos textos dos alunos. Percebe-se que verbos e substantivos compreendem quase metade (47,11%) das palavras utilizadas nos poemas, o que faz sentido, considerando que os poemas têm poucas palavras, e essas classes costumam representar as unidades mais significativas do texto.

Tabela 3: Porcentagens médias de classes gramaticais presentes nos poemas

Classe gramatical	Porcentagem
Verbo	23,56%
Substantivo	23,55%
Pronome	13,08%
Preposição	10,58%
Artigo	10,17%
Conjunção	6,60%
Advérbio	6,51%
Adjetivo	4,57%
Palavra Denotativa	0,92%
Interjeição	0,31%
Numeral	0,19%

5.1.1 Características textuais e séries. Verificando as estruturas que mais costumam ser empregadas por série escolar, destacou-se o caso específico do soneto, uma das formas fixas mais tradicionais. Trata-se de um poema com 14 versos, dispostos em duas quadras seguidas de dois tercetos (na notação aqui utilizada, 4-4-3-3). Conforme apresentado na Tabela 4, observou-se que a quantidade de sonetos tende a aumentar por série, o que parece indicar um interesse crescente em utilizar os recursos comumente empregados no gênero e estudados em sala de aula. Além disso, a utilização dessa estrutura não parece ser uma simples eventualidade, visto que seis desses poemas indicam a escolha pela forma fixa no título, a exemplo do "Soneto de boa noite" [6].

Tabela 4: Distribuição de frequências de sonetos por série

Série	Quantidade de sonetos
9° ano (EF)	1 (0,96%)
1ª série (EM)	3 (2,88%)
2ª série (EM)	4 (4,00%)
3ª série (EM)	6 (6,00%)

Para avaliar a possibilidade de associação entre as demais características textuais (incluindo também os resultados de *PoS tagging*) e a série do(a) autor(a) de cada poema, executou-se um teste τ_b de Kendall para cada uma delas. Na Tabela 5, estão relacionadas as características textuais dos poemas e seus respectivos valores aproximados de τ_b e p-valores, além da indicação da hipótese aceita, considerando-se o nível de significância padrão de 5%.

Percebeu-se, dessa forma, que houve casos em que a hipótese nula H_0 ("a característica não está associada à série do aluno") foi rejeitada, aceitando-se a hipótese alternativa H_1, ou seja, encontrou-se uma associação entre algumas características e as séries dos alunos. A maior parte dessas associações se deu de forma positiva (número de palavras, número de palavras únicas, número de palavras raras, taxa de palavras raras, número de caracteres, média de palavras por verso e porcentagem de interjeições), mas também houve casos de associação negativa (número de estrofes e porcentagem de verbos). Vale mencionar, contudo, que, como o valor de τ_b varia de -1 (associação 100% negativa) a 1 (associação 100% positiva) e as associações que foram avaliadas significativas pelo teste têm valores próximos a zero, essas são consideradas fracas.

Tabela 5: Resultados dos testes τ_b de associação entre características e a série do aluno com nível de significância de 5%

Característica	τ_b	p-valor	Hipótese aceita com significância 5%
Número de palavras	0,058	0,033	H_1
Número de palavras únicas	0,066	0,016	H_1
Número de palavras raras	0,134	0,000	H_1
Diversidade lexical	0,049	0,070	H_0
Taxa de palavras raras	0,129	0,000	H_1
Número de caracteres	0,066	0,015	H_1
Média de tamanho de palavra	-0,001	0,963	H_0
Número de versos	-0,025	0,375	H_0
Média de palavras por verso	0,147	0,000	H_1
Número de estrofes	-0,089	0,003	H_1
% Substantivo	0,034	0,215	H_0
% Verbo	-0,065	0,018	H_1
% Adjetivo	-0,035	0,211	H_0
% Pronome	0,006	0,819	H_0
% Artigo	0,019	0,492	H_0
% Numeral	0,017	0,601	H_0
% Preposição	0,051	0,065	H_0
% Conjunção	0,011	0,679	H_0
% Interjeição	0,105	0,001	H_1
% Advérbio	-0,044	0,112	H_0
% Palavra Denotativa	-0,003	0,929	H_0

Características cujas hipóteses nulas foram rejeitadas estão destacadas em **negrito**.

5.2 Modelagem dos tópicos

Em seguida, aplicou-se o algoritmo de GSDMM, visando agrupar os textos e, dessa forma, determinar os tópicos que os representam. A Tabela 6 ilustra os resultados de *scores* para cinco execuções com cada par de parâmetros α e β, sumarizados por suas médias e desvios padrões. Dessa forma, constatou-se que o par de parâmetros testados que resultou no melhor resultado foi $\alpha = 0{,}5$ e $\beta = 0{,}1$.

Tabela 6: *Scores* dos modelos gerados para cada par de parâmetros α e β

β \ α	**0,01**	**0,1**	**0,25**	**0,5**
0,01	0,049 ± 0,001	0,060 ± 0,006	0,058 ± 0,001	0,062 ± 0,001
0,1	0,070 ± 0,007	0,071 ± 0,011	0,073 ± 0,009	**0,075 ± 0,007**
0,25	0,057 ± 0,008	0,062 ± 0,005	0,062 ± 0,009	0,065 ± 0,007
0,5	0,040 ± 0,007	0,057 ± 0,005	0,042 ± 0,008	0,058 ± 0,005

Melhor resultado destacado em **negrito**.

Dispondo desse resultado, foram realizadas cinco novas execuções com $\alpha = 0{,}5$ e $\beta = 0{,}1$, utilizando mais uma vez o *score* baseado em NPMI como métrica de seleção. Ao final da execução, o melhor modelo selecionado e que será discutido a seguir separou os textos em 44 diferentes grupos, ou seja, determinou 44 tópicos, que podem ser visualizados na íntegra no repositório do projeto[10].

[10] https://github.com/JRobsonJr/pln-poemas/blob/main/dados/topicos.csv

5.2.1 Interpretação dos tópicos. Buscando entender melhor o modelo gerado, foi feita a leitura das dez palavras mais frequentes em cada tópico e tentou-se determinar uma ideia que os resumisse. A fim de complementar o entendimento e avaliar as anotações escolhidas, foi também feita a leitura de pelo menos três poemas de cada tópico, priorizando os que possuíssem o máximo dessas palavras com maior ocorrência. Finalmente, foi feito um recorte do conjunto de tópicos descobertos visando diversificar a discussão, evidenciando noções diferentes percebidas nos resultados. Os tópicos selecionados estão representados na Tabela 7, com suas respectivas quantidades de documentos, rótulos atribuídos e palavras mais frequentes.

Os tópicos T0 e T1, que têm o maior número de representantes, estão relacionados aos temas amor e vida; o primeiro compreende, em especial, ações ("amar", "viver", "fazer", "sei"), e o segundo agrupa termos que expressam ideias absolutas, indicando constâncias ("tudo", "sempre", "nunca", "ninguém"). Esses dois tópicos abarcam noções diversas, indo de lições de vida ("Família é amor / É quem te dá tudo / Sempre te ajudando / A lutar com o mundo." [8]) a declarações de amor ("O que é que faço com o meu coração / se é da minha natureza simplesmente te amar? / Corro o risco de te perder pra sempre." [2]).

Já o tópico T2 remete a uma ideia mais específica: "solidão", "alegria" e "felicidade" são algumas das palavras do tópico que estão no campo dos sentimentos. Um exemplo é o poema "Se sentimentos fossem gente" [23], em que a autora imagina como seria a interação dos sentimentos personificados ("A paixão não seria realidade / A raiva viveria resmungando / E a tristeza, ah, a tristeza...").

As palavras mais comuns do tópico T3 são "poema", "palavras" e "poesia", indicando que esse grupo engloba poemas metalinguísticos, ou seja, que tratam do próprio fazer poético. Esse tópico inclui poemas como o já mencionado "Viagem poética" e também "Brincar de ser poeta" [22] ("Brincando de ser poeta, / mundo, poesia, / palavras, moeda, / junção, alegria.").

É possível perceber que há tópicos associados a ideias similares, como acontece com T13, T29 e T34; todos esses tratam do ambiente em que viviam os autores ao tempo da produção do texto: o sertão paraibano. Por mais que esses tópicos pudessem ser agrupados entre si, percebe-se que há especificidades que os diferenciam:

- T13 trata de uma perspectiva mais pessoal do povo da região (contendo palavras como "povo", "gente", "nordeste", "aqui");
- T29 coloca em perspectiva a seca que acomete o sertão ("sertão", "água", "sede", "seca", "chuva");
- T34 inclui menções a figuras da cultura regional ("baião", "peão", "luiz", "gonzaga").

Há, ainda, especialmente nos grupos com menos representantes, tópicos que capturam ideias específicas, o que não necessariamente corresponde aos seus temas predominantes. T32, por exemplo, traz principalmente termos relacionados com persistência e desistência ("desistir", "conseguir", "podemos", "desistência"), mas também alguns termos próprios do futebol ("bola", "time", "gol"). Isso se manifesta de diferentes formas:

- No poema "Chegou a hora..." [11], a poeta utiliza as imagens do futebol como uma metáfora para a preparação para o vestibular ("Perder, empatar ou ganhar? / O intuito é um sonho realizar / E, assim, esperamos que, ao som do apito final, / Possamos ganhar o jogo da vida real.");
- No poema "Futebol" [1], o poeta utiliza os termos do esporte com o objetivo de homenageá-lo ("O técnico faz de tudo / Para fazer uma escalação / Para no final o time / Se consagrar campeão.").

Quatro dos tópicos descobertos (T40, T41, T42 e T43) foram associados a um único poema cada e, por isso, não foram representados na Tabela 7. Algo em comum entre esses textos é a escolha vocabular, que conta com a presença de palavras com raras ocorrências ou comumente associadas a outros tópicos. Um exemplo é o poema "O corvo" [24], em que o autor denuncia os prejuízos do trabalho

Tabela 7: Exemplos de tópicos

Tópico	Qtd. de poemas	Rótulo atribuído	Palavras mais frequentes
T0	82	Amar e viver	**amor**, **vida**, sempre, **amar**, **coração**, **viver**, **fazer**, sei, tudo, agora
T1	55	Constâncias sobre amor e vida	**tudo**, mundo, **sempre**, **vida**, **amor**, **nunca**, pode, **ninguém**, dia, alegria
T2	49	Sentimentos	vida, **amor**, coração, **solidão**, dia, tempo, pouco, **alegria**, **sentimentos**, **felicidade**
T3	37	Poema e poesia	**poema**, **palavras**, **poesia**, tempo, faz, lá, mundo, fazer, vida, pode
T4	37	Amor	**amor**, sempre, vida, tudo, mundo, sabe, **amar**, **coração**, **alguém**, tempo
T6	27	Infância	**alegria**, **brincar**, **pular**, dia, **criança**, **dançar**, **cantar**, **correr**, bom, viver
T7	27	Realidade brasileira	**mundo**, onde, pessoas, **país**, **nação**, crianças, **futuro**, **brasil**, **pobre**, **desigualdade**
T8	23	Amor e incertezas	**amor**, **vida**, **alma**, **tudo**, **sentido**, sei, ainda, **futuro**, viver, **amar**
T10	21	Sonhos	mundo, deus, **sonhos**, vida, **desistir**, **sonho**, **sonhar**, coração, **acreditar**, nunca
T11	20	Rotina / passagem do tempo	mundo, **tempo**, bem, vida, chão, **dia**, **passar**, **repente**, **porta**, alto
T13	19	Povo nordestino	alegria, coração, **povo**, **nordeste**, **gente**, onde, país, **sol**, **aqui**, **terra**
T15	18	Natureza local	**aqui**, gente, tudo, **chão**, **vida**, **água**, **mundo**, tempo, **ajuda**, melhor
T17	15	Contemplação da natureza	posso, **mar**, onde, bem, **céu**, lá, **luz**, **infinito**, **estrelas**, lugar
T18	12	Amor melancólico	**amor**, **coração**, **dor**, **sentimento**, faz, primeiro, conta, **enche**, vezes, **saudades**
T29	7	Seca no sertão	tudo, **sertão**, **terra**, **água**, sol, **sede**, **seca**, fazer, deus, **chuva**
T32	5	Persistência e desistência	**desistir**, **conseguir**, **podemos**, bola, time, vem, final, gol, nunca, **desistência**
T34	4	Figuras regionais	**rei**, parte, homem, **baião**, **luiz**, **sertão**, **gonzaga**, grandes, **peão**, asa

As palavras mais relacionadas com os rótulos atribuídos estão destacadas em **negrito**.

infantil. Esse texto contém várias palavras que raramente ocorrem em outros poemas, o que pode ter comprometido seu agrupamento. “Curiosidade”, “extensão”, “machado” e “despedida”, por exemplo, são termos que só aparecem novamente em um outro texto cada um, estando associados a tópicos diferentes entre si. Dessa forma, alguns temas podem não ter sido detectados por não haver muita representatividade e/ou pelo emprego de palavras infrequentes.

5.2.2 Temas por série. Na Tabela 8, podem-se ver os principais tópicos por série, representando cada um ao menos 5% dos textos do ano escolar dos autores no momento da escrita.

Tabela 8: Principais temas por série

Série	Temas
6º ano	T1 (14,4%); T6 (10,8%); T13 (8,4%); T0 e T11 (7,2%); T17 (6,0%)
7º ano	T0 (15,2%); T1 (12,1%); T2 e T14 (9,1%); T9 (6,1%)
8º ano	T0 (15,6%); T1 (9,4%); T2 e T7 (7,3%); T3, T4 e T6 (5,2%)
9º ano	T0 e T1 (10,6%); T2 (7,7%); T5 (6,7%); T3 e T4 (5,8%)
1ª série	T0 (13,7%); T2, T5 e T9 (6,9%); T1 (5,9%)
2ª série	T0 (12,0%); T2 e T4 (8,0%); T3, T7 e T8 (7,0%); T12 (6,0%)
3ª série	T3 (12,0%); T0 (9,0%); T2, T4 e T8 (7,0%); T5 e T10 (5,0%)

Percebe-se que amar e viver (T0) costuma ser um dos temas mais presentes, independentemente da série do aluno, assim como sentimentos (T2), presente nos principais temas a partir do 7º ano. Nas fases iniciais, especialmente no caso dos alunos do 6º ano, costuma-se tratar das suas realidades, o que já é conhecido pelos estudantes, como visto pela predominância de tópicos como infância (T6), o povo nordestino (T13), a rotina e a passagem do tempo (T11) e a contemplação da natureza (T17).

Na transição para o ensino médio, percebem-se algumas mudanças: o amor e suas incertezas (T8) torna-se um dos temas mais frequentemente abordados. Outro destaque é o tema relativo a sonhos (T10) aparecendo dentre os mais relevantes dos alunos da 3ª série do ensino médio, o que parece estar relacionado com o período do vestibular e a manifestação dos anseios nessa transição para a universidade e a vida adulta.

Alguns tópicos têm tendências visíveis em relação à série: alunos mais velhos costumam abordar menos tópicos como T1 (constâncias do amor e da vida) e T6 (infância). O amadurecimento e a compreensão das nuances do amor e da vida podem ser uma explicação para a diminuição da quantidade de poemas de T1, assim como a mudança de interesses no período da pré-adolescência e adolescência pode estar associada à presença menor de T6.

Há também tópicos que ficam mais frequentes com o aumento da série: é o caso de T3 e T4. O crescimento de poemas do tópico T3 (poema e poesia) pode estar associado ao maior domínio sobre os instrumentos da produção poética com a formação, enquanto T4, relacionado ao amor (“amor”, “amar”, “coração”, “alguém”), segue a tendência da descoberta crescente desse sentimento nessa faixa etária. As distribuições de poemas dos tópicos T1, T3, T4 e T6 estão ilustradas na Figura 4, agrupadas pela tendência de decrescimento ou crescimento.

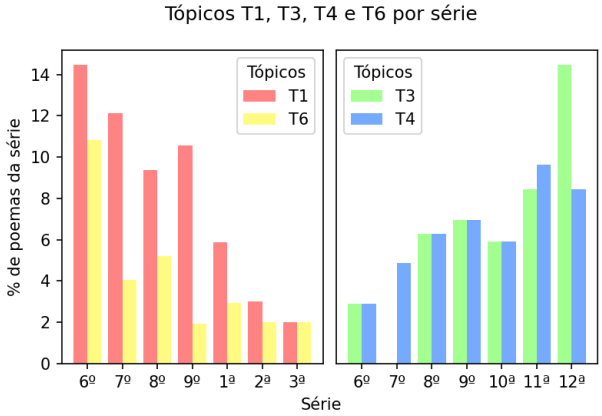

Figura 4: Porcentagens de poemas dos tópicos T1, T3, T4 e T6 por série, agrupados por tendência decrescente e crescente.

5.2.3 Temas por edição. Similarmente à apresentação por série, a Tabela 9 exibe os temas mais frequentes por edição do projeto.

Tabela 9: Principais temas por edição

Edição	Temas
2010	T0 (16,4%); T1 e T3 (6,8%); T2, T5 e T20 (5,5%)
2011	T0 (15,5%); T1 e T5 (11,3%); T4 (7,0%); T2 e T6 (5,6%)
2012	T8 (9,1%); T0 e T2 (7,6%); T4 (6,1%)
2013	T0 (15,8%); T1 (9,2%); T7 (6,6%); T6, T11 e T15 (5,3%)
2014	T0 e T3 (10,3%); T10 (8,8%); T14 (7,4%); T1, T2, T4, T6, T15 e T17 (5,9%)
2015	T0 (11,6%); T3 e T4 (8,7%); T1 (7,2%); T7 e T10 (5,8%)
2016	T0 (15,4%); T2 (10,8%); T1, T3 e T16 (7,7%)
2017	T1 e T2 (11,7%); T0 (8,3%); T7 (6,7%); T5, T17 e T18
2018	T2 (13,2%); T1 (8,8%); T4 e T5 (7,4%); T0, T7 e T9 (5,9%)
2019	T0 (11,8%); T1 (8,8%); T4 (7,4%); T18 (5,9%)

Os tópicos mais comuns no geral, T0 (amar e viver) e T1 (constâncias do amor e da vida), são os que mais aparecem dentre os principais temas por edição; o primeiro aparece em todas, enquanto T1 aparece em nove das dez sob análise. Ainda assim, constatou-se que esses dois e os tópicos T2 (sentimentos), T3 (poema e poesia), T4 (amor), T6 (infância), T7 (realidade brasileira) e T8 (amor e incertezas) estão contemplados em pelo menos um poema de cada edição.

De forma geral, a equipe organizadora do *Coletânea de Poesias* costuma referenciar, na apresentação dos livros, alguns temas que tenham surgido naturalmente a partir dos textos dos alunos. Além disso, o projeto gráfico também busca criar uma conexão entre a capa e os temas representados na edição. A seguir, são apresentadas algumas constatações que relacionam as apresentações dos livros aos temas encontrados:

- na edição de 2010, sugere-se o tema dos sonhos. Percebeu-se T20 (desejos e percepções), dentre os principais dessa edição, como um tópico associado a essa ideia;

- na edição de 2013, são mencionados múltiplos temas, como a alegria de ser criança (relacionado com T6), a indignação pelas injustiças sociais (T7) e os protestos pela preservação da natureza (T15);
- a edição de 2015, também associada à ideia dos sonhos, apresenta T10 (rotulada exatamente de "sonhos") dentre as principais;
- na edição de 2017, coloca-se em evidência o tema de sentimentos (T2), que foi descoberto como o tema predominante dos poemas dessa edição.

Embora os tópicos T13, T29 e T34 (associados à região dos autores no momento da produção dos poemas) não tenham se destacado isoladamente, percebeu-se que, quando agrupados, há um pico evidente na edição de 2012, conforme ilustrado na Figura 5. Esse fato parece estar associado à realidade vivida naquela época: apesar de a seca ser um problema constante no sertão nordestino, o ano de 2012 marcou o início de um período de estiagem prolongada nessa área [18].

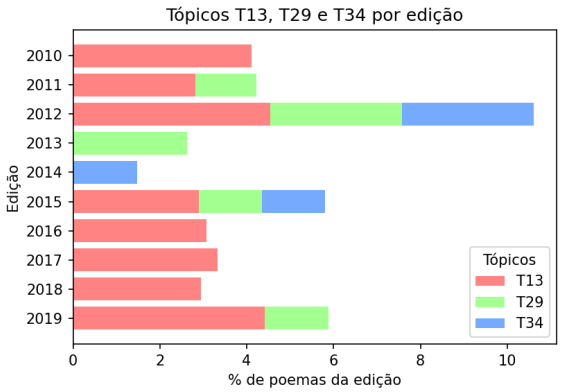

Figura 5: Comparativo por edição da frequência de poemas dos tópicos T13, T29 e T34.

Percebeu-se, ainda, que o tema relativo à infância (T6) não figura entre os principais desde a edição de 2014, aparecendo em, no máximo, 3% dos poemas por edição a partir de 2016. Essa diminuição, embora não muito brusca, parece indicar uma certa mudança geral de interesses, sendo talvez uma marca do amadurecimento cada vez mais precoce.

Em contrapartida, um tema que tem sido colocado em pauta nos anos mais recentes é o T18, rotulado de amor melancólico. Esse tema só começou a ser abordado em 2013, com menor representação, e, desde então, apareceu mais frequentemente, especialmente nas edições de 2017 e 2019. Isso parece indicar uma maior abertura a tratar de temas mais negativos ou íntimos, considerando palavras que figuram nesse tópico (como "dor" e "sofrimento"). A Figura 6 mostra a porcentagem de poemas dos tópicos T6 e T18 por edição.

5.3 Retomando as perguntas de pesquisa

Finalmente, visando entender as implicações dos resultados apresentados ao longo desta seção, em relação às perguntas de pesquisa

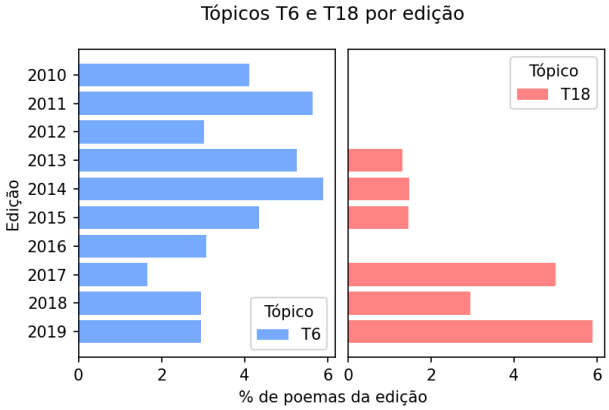

Figura 6: Porcentagens de poemas dos tópicos T6 e T18 por edição.

que nortearam o presente trabalho, retomam-se seus enunciados e apresenta-se uma discussão a respeito de cada uma.

5.3.1 Quais características textuais dos poemas ajudam a diferenciar a etapa de formação (série) dos alunos? Determinou-se, através de testes estatísticos, que as características textuais observadas não têm fortes associações com as séries dos alunos. Essa constatação remete à ideia da liberdade de criação envolvida tanto no projeto, como na própria produção poética. Percebe-se que os alunos utilizam a linguagem de formas diferentes, considerando o que desejam externar através de seus textos. Isso se demonstra também na grande quantidade de formas em que se apresentam os poemas. Ao mesmo tempo, constatações como a do emprego do soneto, ainda que pontualmente, sugerem que há uma atenção e cuidado quanto aos conhecimentos adquiridos. Indica-se, assim, que, à medida que os alunos amadurecem seu contato com o gênero poema, aumenta-se o desejo de aplicar noções já conhecidas nas suas próprias produções.

5.3.2 Como os tópicos abordados pelos alunos mudaram de acordo com a etapa de formação (série)? Percebeu-se que há uma certa mudança de temas associada à faixa etária dos alunos e os seus interesses e vivências. Dessa forma, temas como a infância são menos tratados por alunos mais velhos, enquanto surge a necessidade de expressar sentimentos como o amor e a própria produção poética. Observa-se que há uma constante renovação dos temas, embora a quantidade de tópicos tratados por série não tenha grandes variações.

5.3.3 Como os tópicos abordados pelos alunos mudaram de acordo com os anos (edições do projeto)? Nessa investigação, encontraram-se indícios de que fatos marcantes no ambiente que permeia os estudantes podem ser motivações para suas escritas, como no caso dos textos sobre a seca, na edição de 2012. Ainda, percebe-se uma crescente abertura para tratar de temas com conotação mais íntima, algo que parece indicar positivamente o exercício da sensibilidade por parte dos alunos. Constata-se que as edições têm sempre uma diversidade de temas tratados, o que respalda a intenção do *Coletânea de Poesias* de permitir textos com tema livre.

6 CONCLUSÃO

O Projeto *Coletânea de Poesias* tem se mostrado um exemplo do incentivo ao trabalho com o gênero poema no contexto da educação básica. Os livros confeccionados, frutos materiais desse esforço, contemplam uma riqueza de temas, ao passo que documentam uma parte significativa das trajetórias de alunos que fizeram e fazem parte da história do Fera Colégio e Curso.

Nesta pesquisa, realizou-se a aplicação de técnicas de Processamento de Linguagem Natural para fazer a mineração dos poemas das dez edições mais recentes desse projeto. Através dos resultados obtidos, percebeu-se que as características observadas nos textos não parecem apresentar fortes tendências em relação à série escolar dos estudantes. Em conjunto, identificou-se que não é possível esboçar um único perfil para essas produções textuais. Ambos os aspectos parecem se justificar na liberdade de criação, uma característica indispensável para a produção poética.

Ademais, considerando os resultados da modelagem de tópicos, constatou-se que os alunos estão sujeitos a mudar de temas de acordo com o amadurecimento e a variação de interesses inerente às faixas etárias, além da influência do ambiente que os rodeia. Por fim, também se percebe a influência do contato crescente com o gênero, que parece não somente levá-los a entender melhor as escolhas envolvidas na produção poética, como também ao desejo de aplicá-las em suas próprias escritas. Ao mesmo tempo, constatam-se indícios da aplicabilidade da modelagem de tópicos em poemas e, ainda, da possibilidade de utilizar seus resultados para levar à sala de aula temas dos interesses dos alunos, fomentando o contínuo desenvolvimento do Projeto *Coletânea de Poesias*.

Por outro lado, faz-se necessário observar alguns aspectos que podem consistir em limitações para a presente pesquisa. Como apenas um subconjunto das edições do projeto foi analisado, as tendências constatadas e discutidas neste trabalho podem não representar o *Coletânea de Poesias* em sua integridade. Além disso, percebeu-se que os resultados de modelagem de tópicos também não apresentam temas necessariamente homogêneos, especialmente devido à linguagem plurissignificativa ligada ao gênero poema, sendo que só houve validação dos rótulos atribuídos pela leitura de uma pequena parcela dos textos.

Sugerem-se, assim, trabalhos futuros que possam ajudar a sanar essas possíveis limitações. A utilização das demais edições do projeto pode ajudar a entender o seu panorama completo, além de possivelmente trazer melhores resultados na modelagem de tópicos. Para a validação dessa etapa, ainda, propõe-se a utilização de diferentes técnicas para a sua avaliação automática, bem como para a avaliação manual. Finalmente, visando explorar mais a fundo a produção do *Coletânea de Poesias*, sugere-se a utilização de diferentes técnicas de PLN, a fim de produzir novas caracterizações dos poemas, como a descoberta de rimas e a análise de sentimentos.

AGRADECIMENTOS

Minha gratidão é inversamente proporcional ao espaço que tenho para esta seção, certamente a mais importante deste trabalho. Agradeço então a: Jéssica, por seu amor constante; mãe, por ser minha orientadora de vida; pai, por não poupar esforços pela minha felicidade; minha orientadora, Lívia, por seu acolhimento; Juan e Mariana, meu G3 favorito; Matheus, sempre meu primeiro (co)revisor; e meus demais amigos e familiares, por estarem aqui por mim. Não poderia deixar de agradecer à minha escola do coração, o Colégio GEO Patos, hoje FERA Colégio e Curso, que ainda faz parte de minha trajetória, em nome da diretora Edileny, e à equipe organizadora do *Coletânea de Poesias*, em especial, a Naelma e Vamberlania, praticamente da minha família.

REFERÊNCIAS

[1] João Pedro Severo Alves. 2019. Futebol. In *Coletânea de Poesias*. Vol. 19. FERA Geo, Patos, PB, 17.
[2] Dandhara Tais D. Barros. 2010. O que é que vai ser desse nosso amor? In *Coletânea de Poesias*. Vol. 10. Geo Patos, Patos, PB, 35.
[3] David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*. Chapman and Hall/CRC, 101–124.
[4] Ligia Cademartori. 2012. *O professor e a literatura:* para pequenos, médios e grandes (2 ed.). Autêntica Editora, Belo Horizonte.
[5] Rildo Cosson. 2009. *Letramento literário:* teoria e prática (1 ed.). Editora Contexto, São Paulo.
[6] Bruna de Figueiredo Brito Silva. 2013. Soneto de boa noite. In *Coletânea de Poesias*. Vol. 13. Geo Patos, Patos, PB, 61.
[7] Mateus Clemente de Lacerda. 2015. Identidade nacional. In *Coletânea de Poesias*. Vol. 15. Geo Patos, Patos, PB, 80.
[8] Ludmilla de Menezes Albuquerque Coelho. 2013. Família – a base de tudo. In *Coletânea de Poesias*. Vol. 13. Geo Patos, Patos, PB, 39.
[9] Nikolaos S Dimitriadis. 2020. *Applying Topic Modelling Algorithms on Twitter messages in Greek language*. Graduate Thesis. Aristotle University of Thessaloniki.
[10] Erick R Fonseca, João Luís G Rosa, and Sandra Maria Aluísio. 2015. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. *Journal of the Brazilian Computer Society* 21, 1 (2015), 1–14.
[11] Gabriela Alves Félix. 2011. Chegou a hora... In *Coletânea de Poesias*. Vol. 11. Geo Patos, Patos, PB, 73.
[12] Jorge Junior, Rafael Rossi, and Fabio Lobato. 2019. Uma abordagem baseada em letras para a descoberta de conhecimento da música brasileira: o sertanejo como um estudo de caso. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (Salvador). SBC, Porto Alegre, 949–960.
[13] Daniel Jurafsky and James H. Martin. 2020. Speech and language processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. (2020). Third edition draft.
[14] Harry Khamis. 2008. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography* 24, 3 (2008), 155–162.
[15] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.
[16] Anna Carolyne Gomes Lucena. 2014. Viagem poética. In *Coletânea de Poesias*. Vol. 14. Geo Patos, Patos, PB, 63.
[17] Beth Marcuschi Maria da Graça Costa Val. 2010. Poemas na escola: análise de textos de aluno. *Educação em Revista* 26, 2 (2010), 65–88.
[18] Alexandre Magno Teodosio de Medeiros and Antônio Cavalcanti de Brito. 2017. A seca no Estado da Paraíba–Impactos e ações de resiliência. *Parcerias Estratégicas* 22, 44 (2017), 139–154.
[19] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18, 5 (2011), 544–551.
[20] Borja Navarro-Colorado. 2018. On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. *Frontiers in Digital Humanities* 5 (2018), 15.
[21] Rodolfo Modrigais Strauss Nunes, Ana Carolina Lima de Souza, and Ana Beatriz Bindel Cotait. 2020. PROJETO CONEXÃO LOCAL. (2020).
[22] Izabela Wanderley Nóbrega. 2011. Brincar de poeta. In *Coletânea de Poesias*. Vol. 11. Geo Patos, Patos, PB, 56.
[23] Marta Louise Dantas Dias Oliveira. 2016. Se sentimentos fossem gente. In *Coletânea de Poesias*. Vol. 16. FERA Geo, Patos, PB, 34.
[24] Anderson Candeia Porto. 2013. O corvo. In *Coletânea de Poesias*. Vol. 13. Geo Patos, Patos, PB, 45.
[25] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.
[26] Vera Maria Tietzmann Silva. 2009. *Leitura literária & outras leituras:* impasses e alternativas no trabalho do professor (1 ed.). RHJ, Belo Horizonte.
[27] Talita Sátiro Soares. 2011. Toda família tem. In *Coletânea de Poesias*. Vol. 11. Geo Patos, Patos, PB, 71.
[28] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 233–242.