UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Gabriel Augusto Barbosa

Aprendizagem de Máquina para Classificação de Tipos Textuais: Estudo de Caso em Textos escritos em Português Brasileiro

Recife

2025

Gabriel Augusto Barbosa

Aprendizagem de Máquina para Classificação de Tipos Textuais: Estudo de Caso em Textos escritos em Português Brasileiro

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Péricles Barbosa Cunha de Miranda

Recife

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

B238a Barbosa, Gabriel Augusto.
Aprendisagem de máquinas para classificação de tipos textuais : estudo de caso em textos escritos em português brasileiro / Gabriel Augusto Barbosa. – Recife, 2025.
25 f.: il.

Orientador(a): Péricles Barbosa Cunha de Miranda.
Trabalho de Conclusão de Curso (Graduação) – Universidade Federal Rural de Pernambuco, Departamento de computação, Recife, BR-PE, 2025.
Inclui referências e apêndice(s).

1. Processamento de linguagem natural 2. Classificação textual 3. Tipologia textual 4. Características linguísticas I. Miranda, Péricles Barbosa Cunha de, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Gabriel Augusto Barbosa às 14h do dia 30/07/2025, apresentado remotamente, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado "Aprendizagem de Máquina para Classificação de Tipos Textuais", orientado por Péricles Barbosa Cunha de Miranda e aprovado pela seguinte banca examinadora:

gov.br
Documento assinado digitalmente
PERICLES BARBOSA CUNHA DE MIRANDA
Data: 30/07/2025 23:30:54-0300
Verifique em https://validar.iti.gov.br

ORIENTADOR
DC/UFRPE

gov.br
Documento assinado digitalmente
ANDRE CAMARA ALVES DO NASCIMENTO
Data: 31/07/2025 21:40:24-0300
Verifique em https://validar.iti.gov.br

AVALIADOR
DC/UFRPE

Agradecimentos

Agradeço primeiramente a Deus por me guiar e fortalecer ao longo desta jornada. Aos meus pais, pelo apoio incondicional em todos os momentos. Aos professores que contribuíram para minha formação, em especial Péricles, Rafael e André, por suas orientações e ensinamentos valiosos. Ao laboratório AIboxLab, pelo ambiente de aprendizado e crescimento. Aos amigos que ganhei nesse caminho, com destaque para Hyan e Moésio, cujo apoio e companheirismo foram fundamentais. A todos vocês, minha sincera gratidão — este trabalho só foi possível graças à colaboração e incentivo de cada um.

Resumo

A classificação de textos considerando tipos textuais é de suma importância para algumas aplicações de Processamento de Linguagem Natural (PLN). Nos últimos anos, algoritmos de aprendizado de máquina têm obtido bons resultados nesta tarefa considerando textos em inglês. No entanto, pesquisas voltadas para a detecção de tipos textuais escritos em português ainda são escassas, e ainda há muito a ser estudado e descoberto nesse contexto. Assim, este artigo propõe um estudo experimental que investiga o uso de algoritmos de aprendizado de máquina para classificar textos em português considerando tipos textuais. Para isso, propomos um novo corpus composto por textos em português de dois tipos textuais: narrativo e dissertativo. Três algoritmos de aprendizado de máquina tiveram seu desempenho avaliado no corpus criado em termos de precisão, revocação e pontuação F1. Além disso, também foi realizada uma análise dos atributos envolvidos no processo para identificar quais características textuais são mais importantes na tarefa atual. Os resultados mostraram que é possível alcançar altos níveis de precisão e rememoração na classificação de textos narrativos e dissertativos. Os algoritmos obtiveram níveis de métricas semelhantes, demonstrando a qualidade das características extraídas.

Palavras-chave: PLN, Classificação textual, Tipologia textual, Características linguísticas.

Abstract

The classification of texts regarding textual types is of paramount importance for some Natural Language Processing (NLP) applications. In recent years, machine learning algorithms have achieved good results in this task considering English texts. However, research aimed at detecting textual types written in Portuguese is still scarce, and much remains to be studied and discovered in this context. Thus, this article proposes an experimental study that investigates the use of machine learning algorithms to classify texts in Portuguese regarding textual types. For this, we propose a new corpus composed of Portuguese texts of two textual types: narrative and dissertation. Three machine learning algorithms had their performance evaluated in the proposed corpus in terms of accuracy, recall, and F1 score. Besides, an analysis of the attributes involved in the process was also carried out to identify which textual characteristics are more important in the current task. The results showed that it is possible to achieve high levels of precision and recall in classifying narrative and essay texts. The algorithms obtained similar metrics levels, demonstrating the extracted features' quality.

Keywords: NLP, Text classification, Text typology, Linguistic features.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos últimos anos, pesquisadores de diferentes áreas vem presenciando um aumento significativo na quantidade de textos disponíveis digitalmente (HASSANI et al., 2020), abrindo novos horizontes para diversas aplicações como Análise de Sentimentos (ZHANG; WANG; LIU, 2018) e Tradução de Máquina (DABRE; CHU; KUNCHUKUTTAN, 2020). Em partes, tal aumento se deve a "revolução Big Data"(OUSSOUS et al., 2018) e tendências tecnológicas como a proliferação de dispositivos inteligentes, Internet das Coisas (LI; XU; ZHAO, 2015) e Computação em Nuvem (BOTTA et al., 2016).

Todavia, com um aumento expressivo na quantidade de dados disponíveis, faz-se necessário o uso de diferentes ferramentas para auxiliar a análise e compreensão desses dados (OUSSOUS et al., 2018). Em particular, para textos disponíveis digitalmente, um processo crucial em sua análise é a *classificação* (ZHOU et al., 2020). A classificação de textos consiste em mapear um documento texto a uma ou mais categorias pré-definidas (KOWSARI et al., 2019), que variam de acordo com a aplicação. Por exemplo, é possível classificar textos quanto ao seu domínio, gênero, tipo, e sentimentos (LAGUTINA; LAGUTINA, 2021).

Em específico, a classificação de textos em **tipos** ou **gêneros** textuais é de suma importância para algumas aplicações de Processamento de Linguagem Natural (PLN) (ONAN, 2018). Por exemplo, em sistemas de correção automática de redações, conhecidos como *Automated Essay Scoring* (AES) (KE; NG, 2019), o gênero e tipo do texto são cruciais para avaliação completa de uma redação (PATOUT; CORDY, 2019). Todavia, vide a enorme grande quantidade de textos não classificados disponíveis, a classificação manual se torna inviável e soluções para classificação automática são necessárias (ONAN, 2018; ONAN, 2017).

Na literatura, as definições de gênero e tipo textual são diversas e, por vezes, conflitantes (MELISSOURGOU; FRANTZI, 2017). Para isso, precisamos de uma base teórica de tipologia textual, e aqui serão usadas as definições de "tipelementos" (TRAVAGLIA, 2003). No contexto da organização de categorias textuais, existem 4 "tipelementos", que são: o **tipo**, o **subtipo**, o **gênero** e a **espécie**. Para este trabalho, as classes usadas na classificação serão as definidas sob o contexto de *tipo textual*. Neste contexto, o *tipo* é a classificação mais abrangente de todos os "tipelementos", sendo *espécie*, a classificação mais específica (ou independente) destes. As classes de *Tipo textual* usadas neste trabalho são organizadas em 4 tipos (TRAVAGLIA, 2002):

- **Textos Descritivos**, onde busca-se descrever como é algo;

- **Textos Dissertativos**, que visa refletir, explicar, avaliar ou conceituar algum assunto;
- **Textos Injuntivos**, que tem por objetivo ordenar alguém, ou detalhar a ação requerida ou como fazer;
- **Textos Narrativos**, onde o objetivo é contar uma história, detalhar os acontecimentos.

Ademais, a tipologia textual é de extrema importância no contexto educacional. Por exemplo, o Exame Nacional do Ensino Médio (ENEM) tem como um dos critérios de avaliação a escrita no tipo textual Dissertativo[1]. Além disso, textos narrativos são cruciais para práticas de aprendizagem no ensino fundamental (ROMANO-SOARES; SOARES; CÁRNIO, 2010). Em ambos os casos, os alunos recebem instruções claras do tipo textual a ser desenvolvido e podem até ser eliminados de exames caso não produzam o texto de acordo com o proposto.

Dessa forma, existem diversas estratégias para a classificação automática de tipos textuais na literatura. Por exemplo, (KESSLER; NUNBERG; SCHÜTZE, 1997) demonstrou a capacidade de classificar tipos com base em características chamadas de dicas de superfície (*surface cues*), e ainda criar um detector binário de narratividade. Além disso, vários métodos de classificação de texto têm sido usados no contexto educacional nos últimos anos (FERREIRA-MELLO et al., 2019).

Entretanto, de acordo com o nosso conhecimento, grande parte dos trabalhos possui como ênfase a língua Inglesa, sendo necessário novas pesquisas considerando outros idiomas, principalmente quando se trata da identificação de tipologia textual. Assim, neste trabalho consideramos a língua portuguesa e buscamos responder o seguinte questionamento: “o quão bem textos em português podem ser classificados nessas categorias com base em suas características textuais?”. Para responder a esse questionamento, realizamos um estudo experimental comparando o desempenho de algoritmos de aprendizagem de máquina para a classificação de textos em *tipos textuais*. Inicialmente, devido a ausência de *datasets* em Português para essa tarefa, propomos um corpus com textos classificados em 2 dos 4 tipos textuais mencionados, dissertativo e narrativo, que são os mais utilizados no contexto educacional. Em seguida, extraímos diferentes características dos textos nesse corpus utilizando ferramentas do estado da arte, como o CohMetrix PT-BR (CAMELO; JUSTINO; MELLO, 2020). Por último, comparamos os algoritmos *Random Forest* (RF), *Support Vector Machine* (SVM) e *Stochastic Gradient Descent* (SGD) aplicado à SVM, nesse corpus, em termos de precisão, revocação e pontuação F1.

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

Os resultados sugerem a alta importância dos índices do CohMetrix para este tipo de classificação. Ademais, todos os algoritmos obtiveram resultados relevantes, alcançando valores superiores a 90% em todas as métricas. Por fim, apresentamos as implicações práticas da utilização dos modelos para aplicações educacionais.

2 Trabalhos Relacionados

A classificação de textos é um problema conhecido na área de Processamento de Linguagem Natural (PLN), com várias aplicações em diferentes domínios. Em particular, algoritmos para a classificação automática de textos vem sendo amplamente utilizados para resolver diversos problemas educacionais (FERREIRA-MELLO et al., 2019). Por exemplo, é possível classificar automaticamente mensagens trocadas em fóruns online de ensino a distância (FERREIRA et al., 2018; TEIXEIRA et al., 2020), classificar mensagens de feedbacks fornecidas por professores (CAVALCANTI et al., 2020; CAVALCANTI et al., 2021), ou classificar se redações escritas fogem do tópico proposto (PASSERO et al., 2017). Dessa forma, nessa seção discutimos as diferentes abordagens presentes na literatura para a classificação e análise de textos de acordo com sua tipologia textual.

Os autores em (KESSLER; NUNBERG; SCHÜTZE, 1997) usaram redes neurais para a classificação de tipos textuais, mais especificamente na detecção de narratividade. Para extrair os atributos preditores do texto, o conceito de “dicas genéricas” (*Generic Cues*) foi introduzido, extraindo quatro tipos de “dicas”: dicas estruturais (*Structural Cues*), que utilizam informações como POS (*part-of-speech*); dicas lexicais, que se referem a vocabulários com campos lexicais específicos facilitando a classificação; dicas a nível de letra, usando majoritariamente pontuações; dicas derivadas das anteriores como proporções também foram usadas.

Em (STAMATATOS; FAKOTAKIS; KOKKINAKIS, 2000), os autores usam a frequência de palavras como preditor de tipos textuais. Desta forma, é necessário que os textos da mesma categoria apresentem estilos parecidos, ou seja, textos de mesma classe devem apresentar baixa variação. Por isto, o corpus do *Wall Street Journal* foi usado e foram considerados os tipos: *Editorials*; *Letters to the Editor*; *Reportage*; *Spot news*. No artigo, foram usados trechos de tamanho homogêneo, mas a forma de identificar as categorias se diferencia por causa do uso da análise discriminante, assim como em (KARLGREN; CUTTING, 1994).

Os autores em (BALINT; DASCALU; TRAUSAN-MATU, 2016) também usaram análise de discriminante para a classificação de gêneros textuais. No entanto, diferente dos trabalhos anteriores, considerou características rítmicas, para detecção dos seguintes gêneros: *artigo*, *redação* e *discurso*. Foram usados os corpora *Speeches*, *RST-DT* e *Uppsala Student English* para para fazer os experimentos, todos os corpora usados estão, assim como trabalhos anteriormente citados, na língua inglesa.

Em (ONAN, 2018), o autor considera uma classificação textual baseada na

análise da função da linguagem (WACHSMUTH; BUJNA, 2011), do inglês *language function analysis* ou LFA, nessa teoria os textos são divididos em 3 de gêneros textuais: expressivo, apelativo e informativo. E destes gêneros, os autores propõem o *LFA-corpus* que é composto de textos de *reviews* de livros e câmeras em inglês. Foram usados múltiplos modelos de aprendizado de máquina e obtendo resultados de até 94,3% de acurácia.

Apesar de todos os trabalhos apresentados focarem na extração de tipologia textual, ainda existem limitações claras na literatura, são elas: (i) não foram encontrados trabalhos que tratam de textos em português; (ii) os trabalhos propostos não fazem relação direta com problemas educacionais; (iii) as características utilizadas nos trabalhos anteriores são diversas, mas não foram aplicadas características linguísticas que demonstram boa performance para avaliação de textos educacionais (LAGUTINA; LAGUTINA, 2021). Além disso, não existe um banco de dados consolidado para análise de tipologia textual em português. O presente trabalho visa contribuir na solução de tais limitações.

3 Materiais e Métodos

Nesta seção são apresentados a base de dados (corpus) proposta, as características consideradas para o processo de classificação, os algoritmos de classificação avaliados e as métricas de avaliação adotadas. Os experimentos foram executados em uma máquina com um Intel® Core™ i5-8265U, com 12GB de memória ram e gráficos integrados, a linguagem de programação usada foi Python, com o auxílio das bibliotecas: scikit-learn, NumPy, pandas e cohmetrixBR.

3.1 Base de Dados

Para fazer a classificação de tipos textuais, as características precisam ser extraídas de um corpus textual anotado com as categorias utilizadas. Neste trabalho foi criado um novo corpus textual para a classificação textual, denominado **Corpus de Tipos Textuais Brasileiros** (Corpus TTBR). No momento deste artigo, o corpus proposto contém 2 dos quatro tipos textuais definidos por (TRAVAGLIA, 2018), os tipos narrativo e dissertativo. Estes tipos de texto foram escolhidos devido a sua importância no contexto educacional. Os textos dissertativos foram obtidos da base de dados *uol-redacoes* [1], já os textos narrativos foram obtidos do corpus **OBras** [2].

O TTBR foi criado para a avaliação de algoritmos de aprendizagem de máquina na tarefa de classificação de tipos textuais. Por ser composto de tipos textuais e não gêneros textuais, é esperado que as classificações sejam estáveis ao longo do tempo e a variabilidade de cada tipo seja grande. Ou seja, dentro de cada tipo textual há muitas formas de escrita, por exemplo, nos textos narrativos, existem inúmeros gêneros como: romances, poesias, contos, e outros.

A Tabela 1 mostra a composição do TTBR em termos de número de textos e quantidade média de palavras e caracteres. Como se pode ver, o TTBR é balanceado, possuindo 2164 textos de cada tipo.

Tabela 1 – Composição do Corpus TTBR após o balanceamento.

Tipo Textual	**Textos**	**Caracteres**		**Palavras**	
		Média	Desvio	Média	Desvio
Dissertativo	2164	1568.51	467.45	293.68	87.53
Narrativo	2164	1117.88	570.11	240.14	111.17

[1] https://github.com/gpassero/uol-redacoes-xml
[2] https://www.linguateca.pt/acesso/corpus.php?corpus=OBRAS

Vale salientar que a seleção dos textos narrativos foi feita através da seleção aleatória das obras disponíveis no corpus *OBras*, que incluem obras de Aluízio Azevedo, Machado de Assis, dentre outros autores. Por se tratarem de livros e livretos, os textos do corpus *OBras* possuem grande variação em seu tamanho e forma. Para isso os trechos selecionados foram pré-processados para preservar a estrutura original, filtrando pela categoria prosa e reduzindo a diferença de tamanhos entre eles limitando-os a 24 linhas de cada amostra.

3.2 Características textuais

A classificação de textos só pode ser feita através da extração de características textuais. Neste trabalho, as características foram extraídas através do uso da ferramenta CohMetrix PT-BR (CAMELO; JUSTINO; MELLO, 2020), uma versão em português brasileiro do Coh-Metrix (MCNAMARA et al., 2014). O CohMetrix é uma ferramenta de extração de características (ou índices) textuais para análise de textos das mais variadas fontes, como artigos, redações, instruções e respostas de questionários (CAMELO; JUSTINO; MELLO, 2020). Essa ferramenta foi utilizada devido a sua grande eficácia de análise de textos educacionais(FERREIRA-MELLO et al., 2019). A ferramenta possui uma série de extratores, que são descritos a seguir:

- **Descritivos**: Iniciados por DES, extraem informações descritivas do texto como quantidade, comprimento e variabilidade de parágrafos e sentenças;
- **Coesão Referencial**: Iniciados por CRF, extraem informações referentes a coesão e sobreposição de palavras entre as sentenças adjacentes;
- ***Latent Semantic Analysis***: Iniciados por LSA, medem o nível de sobreposição semântica entre sentenças e parágrafos usando LSA;
- **Diversidade Léxica**: Iniciados por LD, extraem características que calculam informações relacionadas ao vocabulário, como a quantidade de palavras únicas no texto,
- **Conectivos**: Iniciados por CNC, extraem características que medem o número de conectivos no texto, conectivos são frases que conectam sentenças, como: “portanto” e “logo que”.
- **Modelo Situacional**: Iniciados por SM, extraem características que medem o nível de representação mental do texto.
- **Complexidade Sintática**: Iniciados por SYN, extraem informações relacionadas as informações de *part-of-speech* das palavras do texto, criando arvores sintáticas para a avaliação de suas complexidades.

- **Densidade de Padrões Sintáticos**: Iniciados por DR, tratando também da sintaxe, identificam frequências de padrões sintáticos a nível de frase, como frases verbais ou nominais;
- **Informação da Palavra**: Iniciados por WRD, extraem características relacionadas à frequência dos tipos de palavras, como por exemplo: substantivos, verbos, adjetivos e pronomes;
- **Legibilidade**: Iniciados por RD, extraem características que medem o nível de facilidade da leitura do texto.

Neste trabalho, utilizamos tais características como atributos para a classificação de tipos textuais. Os extratores de características descritivas foram desconsiderados, pois estes utilizam informações que dependem da forma como os textos foram obtidos e se houve algum pré-processamento. Visto que os textos do corpus TTBR passaram por uma etapa de pré-processamento e balanceamento, tais índices poderiam refletir erroneamente características dos textos. Por exemplo, no TTBR os textos dissertativos possuem uma média de 4,38 linhas de texto por amostra com uma única quebra de linha, enquanto os textos narrativos possuem uma formatação distinta e por vezes possuem mais de uma quebra de linha por amostra. Dessa forma, a Tabela 2 lista todos os índices considerados neste trabalho. Mais informações sobre as características podem ser encontradas em (CAMELO; JUSTINO; MELLO, 2020).

Tabela 2 – 75 Índices Linguístico Disponíveis

CNCADC	CNCNeg	DRGERUND	SMINTEp	WRDFRQa
CNCAdd	CNCPos	DRINF	SMINTEp_sentence	WRDFRQc
CNCAll	CNCProp	DRNEG	SMINTEr	WRDFRQmc
CNCAlter	CNCTemp	DRNP	SYNLE	WRDIMGc
CNCCaus	CRFAO1	DRPP	SYNMEDlem	WRDMEAc
CNCComp	CRFAOa	DRPVAL	SYNMEDpos	WRDNOUN
CNCConce	CRFCWO1	DRVP	SYNMEDwrd	WRDPRO
CNCConclu	CRFCWO1d	LDMTLDa	SYNNP	WRDPRP1p
CNCCondi	CRFCWOa	LDTTRa	SYNSTRUTa	WRDPRP1s
CNCConfor	CRFCWOad	LDTTRc	SYNSTRUTt	WRDPRP2
CNCConse	CRFNO1	LDVOCDa	WRDADJ	WRDPRP2p
CNCExpli	CRFNOa	RDFKGL	WRDADV	WRDPRP2s
CNCFinal	CRFSO1	RDFRE	WRDAOAc	WRDPRP3p
CNCInte	CRFSOa	RDL2	WRDCNCc	WRDPRP3s
CNCLogic	DRAP	SMCAUSwn	WRDFAMc	WRDVERB

3.3 Algoritmos de Classificação

Foram selecionados três algoritmos de aprendizagem de máquina, popularmente utilizados para tarefas de classificação: Floresta Aleatória (BREIMAN, 2001), do inglês *Random Forest* (RF), Máquina de vetores de suporte, do inglês *Support Vector Machine* (SVM) (AWAD; KHANNA, 2015) e por fim o classificador de gradiente descendente estocástico, do inglês *Stochastic Gradient Descent* (SGD), que se trata do algoritmo de otimização usado em uma SVM. Tais algoritmos foram escolhidos pelos bons resultados obtidos em estudos anteriores (FERNÁNDEZ-DELGADO et al., 2014). Foram utilizadas as parametrizações padrão da biblioteca scikit-learn [3] para cada um dos algoritmos.

Antes da execução dos experimentos, realizamos uma seleção das 75 características para identificar a relevância de cada uma delas para o problema. Existem diferentes métodos para a seleção de características (CHANDRASHEKAR; SAHIN, 2014). Neste trabalho, utilizamos a seleção baseada na permutação proposta em (ALTMANN et al., 2010), utilizando o valor 30 como semente de geração aleatória e o modelo de floresta aleatória para cálculo da importância.

A avaliação dos algoritmos foi feita usando métricas clássicas para classificação (HOSSIN; SULAIMAN, 2015): Precisão, Revocação e F1. Tais métricas são definidas através dos seguintes conceitos de classificação: *TP = True Positive* (Predição certa para caso verdadeiro), *TN = True Negative* (Predição certa para caso falso), *FP = False Positive* (Predição errada para caso verdadeiro), *FP = False Negative* (predição errada para caso falso). Assim, cada uma das métricas é calculada da seguinte forma:

$$\textit{Precisão} = \frac{TP}{TP + FP}, \tag{3.1}$$

$$\textit{Revocação} = \frac{TP}{TP + FN}, \tag{3.2}$$

$$\textit{Pontuação F1} = \frac{2 * \textit{Precisão} * \textit{Revocação}}{\textit{Precisão} + \textit{Revocação}}. \tag{3.3}$$

Como se pode ver, a precisão calcula o percentual de acertos da classe positiva diante de todas as predições positivas. A revocação (ou cobertura), indica o quanto o modelo está identificando os casos positivos corretamente. Por fim, a pontuação F1 calcula a média harmônica entre a precisão e a revocação.

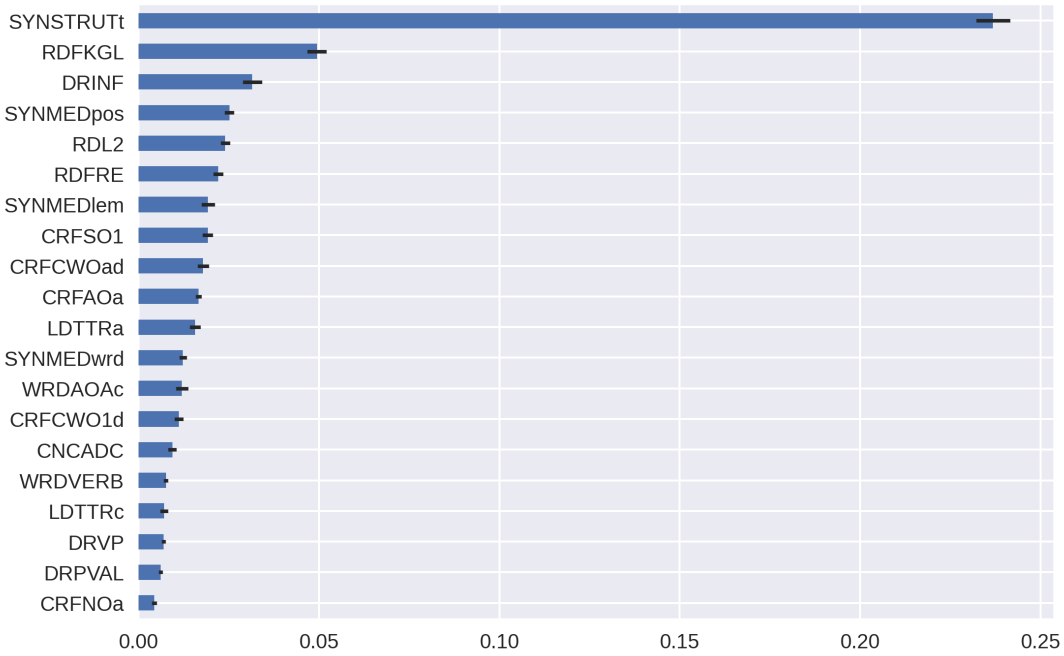
[3] https://scikit-learn.org/stable/

4 Resultados

4.1 Seleção de Características

Como mencionado na Capítulo 3, a primeira etapa do experimento consistiu em selecionar as características mais importantes. A Figura 1 apresenta os valores médio e de desvio padrão da importância dos 20 índices textuais mais importantes.

Figura 1 – Importância dos 20 melhores *índices* por permutação



Os resultados mostram que o índice *SYNSTRUTt* apresentou a maior importância para a classificação textual. O SYNSTRUTt calcula a semelhança entre as árvores sintáticas de duas sentenças, obtendo a média de todas as combinações de árvores de um texto, ou seja, aponta o quanto a estrutura de um texto é uniforme entre si (Equação 4.1).

$$SYNSTRUTt = \frac{\textit{nós em comum}}{\textit{total de nós - nós em comum}} \tag{4.1}$$

A elevada importância da SYNSTRUTt pode se dar graças a semelhança sintática entre sentenças nos textos argumentativos, enquanto os textos narrativos possuem maior variação na sua sintaxe, podendo muitas vezes conter falas de personagens, múltiplos tipos de narração e descrição do cenário.

4.2 Análise dos Modelos de Classificação

A Tabela 3 apresenta os resultados médios obtidos por cada um dos algoritmos em termos de precisão, revocação e pontuação F1. Os algoritmos foram avaliados em 3 cenários distintos: (i) utilizando o SYNSTRUTt como única característica para classificação textual; (ii) usando as 20 características de maior importância; e (iii) usando todas as 75 características.

Os resultados mostram que os modelos, de forma geral, obtiveram bom desempenho na detecção dos tipos textuais. Entretanto, os modelos treinados utilizando apenas o índice SYNSTRUTt alcançaram resultados superiores aos seus pares quando consideramos as métricas revocação e pontuação F1.

No modelo de SVM, o valor de precisão médio com o índice SYNSTRUTt foi de 96,51%, superando de 4% a 3% o mesmo modelo com 20 e 75 índices textuais. As métricas de revocação demonstra uma melhora de 10% entre o modelo com apenas a SYNSTRUTt e o modelo com 20 características. Tanto o SVM quanto o modelo SGD, possuem um desempenho melhor em todas as métricas de avaliação quando usando apenas o índice SYNSTRUTt do Cohmetrix. O que demonstra um alto nível de correlação entre a complexidade sintática e a tipologia textual.

Tabela 3 – Comparação de modelos com 1, 20 e 75 índices do CohMetrix

		SYNSTRUTt		**20 features**		**75 features**	
		Média	Desvio	Média	Desvio	Média	Desvio
RF	Precisão	0.9708	0.0098	**0.9914**	0.0051	0.9861	0.0069
	Revocação	**0.9787**	0.0081	0.9557	0.0211	0.9205	0.0278
	Pontuação F1	**0.9747**	0.0066	0.9731	0.0107	0.9520	0.0159
SVM	Precisão	**0.9651**	0.0106	0.9277	0.0147	0.9332	0.0182
	Revocação	**0.9945**	0.0045	0.8956	0.0300	0.8572	0.0300
	Pontuação F1	**0.9795**	0.0059	0.9112	0.0197	0.8934	0.0209
SGD	Precisão	**0.9643**	0.0122	0.9439	0.0495	0.9557	0.0257
	Revocação	**0.9935**	0.0047	0.8803	0.1168	0.9275	0.0843
	Pontuação F1	**0.9786**	0.0068	0.9037	0.0527	0.9383	0.0419

No geral os três modelos obtiveram ótimos resultados na classificação dos tipos textuais no TTBR, com métricas de precisão e revocação acima de 95% no melhor caso. O índice SYNSTRUTt (CAMELO; JUSTINO; MELLO, 2020), como sugere a Figura 1, se mostrou como o mais relevante para a classificação. Todavia, a alta precisão, superior a 99,1%, e baixo desvio padrão, de 0,5%, do algoritmo RF com 20 características sugere que outros índices textuais contribuem significativamente para a predição, sendo estes os melhores resultados, em precisão, para todos algoritmos testados no TTBR.

Por fim, vale destacar que o algoritmo RF conseguiu os melhores resultados

quando consideradas outras características além do índice SYNSTRUTt. Conjecturamos que isso ocorreu por esse algoritmo realizar um processo interno de seleção de características em suas etapas de processamento, permitindo uma reorganização das características mais importantes. Dessa forma, o RF apresentou melhores resultados quando consideradas múltiplas características.

4.3 Implicações práticas na área educacional

Como já apresentado anteriormente, a identificação de tipologia textual é uma atividade extremamente relevante para a aplicações educacionais, todavia, estudos neste tópico possuem poucos trabalhos considerando textos em português brasileiro. Com os modelos criados aqui, é possível auxiliar professores na correção de um dos critérios do ENEM, assim como a análise de textos de alunos do ensino fundamental (SILVA, 2012).

É importante destacar que os classificadores propostos neste trabalho podem servir como base para o desenvolvimento de ferramentas de suporte ao professor na correção de produções textuais de alunos de diferentes níveis. Este trabalho está dentro do contexto de um projeto maior que tem como objetivo a recuperação de aprendizado de alunos de escolas públicas[1]. Além disto, a abordagem proposta também pode ser facilmente adaptada para outros critérios de de avaliação de produções textuais, como para a análise de coesão (LAPATA; BARZILAY, 2005), já que estes também consideram diferentes elementos textuais.

[1] Mais detalhes sobre o projeto foram omitidos para respeitar a revisão às cegas e serão incluídos na versão final.

5 Conclusões

Este trabalho apresentou um novo corpus, denominado Corpus TTBR, para classificação de tipos textuais em português trazendo inicialmente textos dissertativos e narrativos. Além disso, foram selecionadas as características do CohMetrix com maior importância para o problema de classificação de tipos textuais, onde o índice SYNSTRUTt se demonstrou crucial para esse problema. Por fim, foi realizada uma análise experimental, usando algoritmos aprendizado de máquina clássicos, para a classificação dos tipos textuais presentes no TTBR. Os resultados demonstraram alta performance de todos algoritmos em todos os testes.

A maior limitação do presente trabalho é o número limitado de exemplos de textos dissertativos disponíveis para composição do TTBR, fazendo com que fosse necessário reduzir a quantidade de textos narrativos na composição do dataset para mantê-lo balanceado. Ademais, a falta de textos de outros tipos textuais, como descritivos e injuntivos, pode limitar a aplicabilidade dos modelos treinados.

Dessa forma, trabalhos futuros podem incrementar o corpus TTBR com novos tipos textuais e realizar novos experimentos para validar os resultados obtidos neste trabalho. Por último, os modelos criados podem ser utilizados para auxiliar a sugestão de notas e feedback de professores em produções textuais de alunos.

Referências

ALTMANN, A. et al. Permutation importance: a corrected feature importance measure. *Bioinformatics*, Oxford University Press, v. 26, n. 10, p. 1340–1347, 2010. Citado na página 16.

AWAD, M.; KHANNA, R. Support vector machines for classification. In: *Efficient learning machines*. [S.l.]: Springer, 2015. p. 39–66. Citado na página 16.

BALINT, M.; DASCALU, M.; TRAUSAN-MATU, S. Classifying written texts through rhythmic features. In: SPRINGER. *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. [S.l.], 2016. p. 121–129. Citado na página 11.

BOTTA, A. et al. Integration of cloud computing and internet of things: A survey. *Future Generation Computer Systems*, v. 56, p. 684–700, 2016. ISSN 0167-739X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X15003015>. Citado na página 8.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 16.

CAMELO, R.; JUSTINO, S.; MELLO, R. F. L. de. Coh-metrix pt-br: Uma api web de análise textual para a educação. In: SBC. *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*. [S.l.], 2020. p. 179–186. Citado 4 vezes nas páginas 9, 14, 15 e 18.

CAVALCANTI, A. P. et al. Utilização de recursos linguísticos para classificação automática de mensagens de feedback. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 861–872. Citado na página 11.

CAVALCANTI, A. P. et al. Análise automática de feedback em ambientes de aprendizagem online. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 892–901. Citado na página 11.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 1, p. 16–28, 2014. Citado na página 16.

DABRE, R.; CHU, C.; KUNCHUKUTTAN, A. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 53, n. 5, sep 2020. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3406095>. Citado na página 8.

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014. Citado na página 16.

FERREIRA, M. A. D. et al. Um sistema baseado em pln e ag para apoiar a mediação pedagógica em fóruns de discussão. *Revista Brasileira de Informática na Educação*, v. 26, n. 03, p. 61, 2018. Citado na página 11.

FERREIRA-MELLO, R. et al. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 9, n. 6, p. e1332, 2019. Citado 3 vezes nas páginas 9, 11 e 14.

HASSANI, H. et al. Text mining in big data analytics. *Big Data and Cognitive Computing*, v. 4, n. 1, 2020. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/4/1/1>. Citado na página 8.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015. Citado na página 16.

KARLGREN, J.; CUTTING, D. *Recognizing Text Genres with Simple Metrics Using Discriminant Analysis*. arXiv, 1994. Disponível em: <https://arxiv.org/abs/cmp-lg/9410008>. Citado na página 11.

KE, Z.; NG, V. Automated essay scoring: A survey of the state of the art. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2019. p. 6300–6308. Disponível em: <https://doi.org/10.24963/ijcai.2019/879>. Citado na página 8.

KESSLER, B.; NUNBERG, G.; SCHÜTZE, H. Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*, 1997. Citado 2 vezes nas páginas 9 e 11.

KOWSARI, K. et al. Text classification algorithms: A survey. *Information*, v. 10, n. 4, 2019. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/10/4/150>. Citado na página 8.

LAGUTINA, K.; LAGUTINA, N. A survey of models for constructing text features to classify texts in natural language. In: *2021 29th Conference of Open Innovations Association (FRUCT)*. [S.l.: s.n.], 2021. p. 222–233. ISSN 2305-7254. Citado 2 vezes nas páginas 8 e 12.

LAPATA, M.; BARZILAY, R. Automatic evaluation of text coherence: Models and representations. In: *IJCAI*. [s.n.], 2005. p. 1085–1090. Disponível em: <http://ijcai.org/Proceedings/05/Papers/0505.pdf>. Citado na página 19.

LI, S.; XU, L. D.; ZHAO, S. The internet of things: a survey. *Information Systems Frontiers*, v. 17, n. 2, p. 243–259, Apr 2015. ISSN 1572-9419. Disponível em: <https://doi.org/10.1007/s10796-014-9492-7>. Citado na página 8.

MCNAMARA, D. S. et al. *Automated evaluation of text and discourse with Coh-Metrix*. [S.l.]: Cambridge University Press, 2014. Citado na página 14.

MELISSOURGOU, M. N.; FRANTZI, K. T. Genre identification based on sfl principles: The representation of text types and genres in english language teaching material. *Corpus Pragmatics*, v. 1, n. 4, p. 373–392, Dec 2017. ISSN 2509-9515. Disponível em: <https://doi.org/10.1007/s41701-017-0013-z>. Citado na página 8.

ONAN, A. Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*, Emerald Publishing Limited, v. 46, n. 2, p. 330–348, Jan 2017. ISSN 0368-492X. Disponível em: <https://doi.org/10.1108/K-10-2016-0300>. Citado na página 8.

ONAN, A. An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, SAGE Publications Sage UK: London, England, v. 44, n. 1, p. 28–47, 2018. Citado 2 vezes nas páginas 8 e 11.

OUSSOUS, A. et al. Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, v. 30, n. 4, p. 431–448, 2018. ISSN 1319-1578. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1319157817300034>. Citado na página 8.

PASSERO, G. et al. Off-topic essay detection: A systematic review. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 51. Citado na página 11.

PATOUT, P.-A.; CORDY, M. Towards context-aware automated writing evaluation systems. In: *Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence*. New York, NY, USA: Association for Computing Machinery, 2019. (EASEAI 2019), p. 17–20. ISBN 9781450368520. Disponível em: <https://doi.org/10.1145/3340435.3342722>. Citado na página 8.

ROMANO-SOARES, S.; SOARES, A. J. C.; CÁRNIO, M. S. Práticas de narrativas escritas em estudantes do ensino fundamental. *Pró-Fono Revista de Atualização Científica*, SciELO Brasil, v. 22, p. 379–384, 2010. Citado na página 9.

SILVA, P. N. d. *Tipologias textuais: como classificar textos e sequências*. [S.l.]: CELGA/Livraria Almedina, 2012. Citado na página 19.

STAMATATOS, E.; FAKOTAKIS, N.; KOKKINAKIS, G. Text genre detection using common word frequencies. In: *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*. [S.l.: s.n.], 2000. Citado na página 11.

TEIXEIRA, J. B. et al. Classificação automática da presença social em discussões online escritas em português. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 942–951. Citado na página 11.

TRAVAGLIA, L. C. Tipos, gêneros e subtipos textuais e o ensino de língua materna. *Língua Portuguesa: uma visão em mosaico. São Paulo: EDUC*, p. 201–214, 2002. Citado na página 8.

TRAVAGLIA, L. C. Tipelementos e a construção de uma teoria tipológica geral de textos. *FÁVERO, Leonor Lopes; BASTOS, Neusa M. de O. Barbosa*, p. 97–117, 2003. Citado na página 8.

TRAVAGLIA, L. C. Tipologia textual e ensino da língua. *A ser publicado como capítulo do livro Linguística Textual e Análise da conversação (GTLAC) da ANPOLL. Uberlândia*, 2018. Citado na página 13.

WACHSMUTH, H.; BUJNA, K. Back to the roots of genres: Text classification by language function. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. [S.l.: s.n.], 2011. p. 632–640. Citado na página 12.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, v. 8, n. 4, p. e1253, 2018. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253>. Citado na página 8.

ZHOU, X. et al. A survey on text classification and its applications. *Web Intelligence*, IOS Press, v. 18, p. 205–216, 2020. ISSN 2405-6464. 3. Disponível em: <https://doi.org/10.3233/WEB-200442>. Citado na página 8.