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

Gabriel A. Barbosa[1], **Hyan H. N. Batista**[1], **Péricles Miranda**[1], **Jário Santo**[4], **Seiji Isotani**[4,5], **Thiago Cordeiro**[3], **Ig Ibert Bittencourt**[3,5], **Rafael Ferreira Mello**[1,2]

[1] Universidade Federal Rural de Pernambuco

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)

[3]Universidade Federal de Alagoas

[4]Universidade de São Paulo

[5]Harvard Graduate School of Education

{gabriel.a07b,hyanbatista42,jario.infor}@gmail.com

{pericles.miranda,rafael.mello}@ufrpe.br

thiago@ic.ufal.br,{seiji_isotani,ig_bittencourt}@gse.harvard.edu

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.*

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*

DOI: 10.5753/sbie.2022.224769

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.

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 et al. 2018] e Tradução de Máquina [Dabre et al. 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 et al. 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, autor, gênero, tipo, sentimentos, entre outros [Lagutina and 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 and Ng 2019, Ferreira Mello et al. 2022], o gênero e tipo do texto são cruciais para avaliação completa de uma redação ou questão discursiva [Patout and Cordy 2019, Ferreira-Mello et al. 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 and 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: o tipo **descritivo**, o tipo **dissertativo**, o tipo **injuntivo** e o tipo **narrativo** [Travaglia 2018]. Em resumo, os tipos textuais [Travaglia 2002] podem ser definidos como:

- **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 dizer a ação requerida ou como fazer;
- **Textos Narrativos**: onde o objetivo é contar uma história, dizer os acontecimentos.

Existem diversas estratégias para a classificação automática de tipos textuais na literatura. Por exemplo, [Kessler et al. 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.

Entretanto, de acordo com o nosso conhecimento, grande parte dos trabalhos possui como ênfase a língua Inglesa, sendo necessário novas pesquisas considerando outros idiomas. 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, vide a ausência de *datasets* em Português para essa tarefa, propomos um corpus com textos classificados em 2 dos 4 tipos textuais mencionados. Em seguida, extraímos diferentes características dos textos nesse corpus através da ferramenta CohMetrix PT-BR [Camelo et al. 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.

Os resultados sugerem a alta importância de um dos índices do CohMetrix para este tipo de classificação. Além disso, todos os algoritmos obtiveram ótimos resultados, alcançando valores superiores a 90% em todas as métricas.

O restante do artigo é estruturado da seguinte forma: A Seção 2 discute os trabalhos relacionados à classificação de textos em tipos e gêneros textuais. As seções 3 e 4 apresentam, respectivamente, o arranjo experimental e os resultados obtidos. Por último, a Seção 5 resume o artigo e discute trabalhos futuros.

2. Trabalhos Relacionados

A classificação de textos é uma tarefa conhecida na área de PLN, principalmente em inglês. Em em [Karlgren and Cutting 1994], os autores criaram uma forma simplificada de fazer tal classificação usando características textuais simples como contagens de certas palavras. O artigo utiliza o corpus *Brown* para os experimentos, separando em 2, 4 e 15 categorias, o método de classificação textual utilizou análise discriminante [Mustonen 1965].

Os autores em [Kessler et al. 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 et al. 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 and Cutting 1994].

Os autores em [Balint et al. 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 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 and 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.

Um aspecto comum observado nos artigos anteriores sobre o tema, é que não há um *dataset* de tipos textuais gerais no português. São sempre textos categorizados com classes selecionadas a partir da fonte do texto. Como definido por [Kessler et al. 1997], tipo textual possui somente o significado de definir a origem ou propósito do texto, ex: jornal, receita. Mas sem uma categorização bem definida de tipologia num contexto geral.

Dentre as várias formas de classificação textual, o presente trabalho utiliza uma abordagem de classificação com uso de características linguísticas, como demonstrado por trabalhos citados por [Lagutina and Lagutina 2021]. Ao todo foram considerados 4 tipos textuais baseados em características linguísticas: descritivo, dissertativo, injuntivo e narrativo. O presente artigo avalia diferentes algoritmos de aprendizagem de máquina na classificação de textos em Português levando em consideração 2 dos 4 tipos de textuais mencionados (tipos narrativo e dissertativo). Devido a ausência de bases de dados em Português para esta temática, foi criada a base dados chamada *Corpus de Tipos Textuais Brasileiros* (TTBR) a fim de avaliar os algoritmos selecionados.

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. 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. Corpus TTBR após o balanceamento

	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

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 et al. 2020], uma versão em português brasileiro do Coh-Metrix produzido por [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 et al. 2020]. 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;

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

- ***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.

No contexto deste trabalho, tais características são usadas como atributos para a classificação de tipos textuais. Extratores de características descritivas foram desconsiderados, pois este tipo de características textuais possui informação que têm dependência em como os textos foram obtidos e se houve pré-processamento. Por exemplo, o índice DESSC que revela o número de sentenças ou DESWC que conta o número de palavras podem descrever diferenças de tamanho, que podem ter sido alteradas na obtenção do texto. Outro bom exemplo é o número de parágrafos. Como os textos dissertativos do TTBR possuem uma estrutura de texto mais simples, tendo uma média de 4,38 linhas de texto por amostra e considerando parágrafos como uma única quebra de linha, isso o diferencia claramente do texto narrativo, onde os parágrafos podem ter sido idealizados como quebras duplas de linha. Isto causa uma diferença de formatação que pode causar um *overfiting* nos classificadores, por isso estes atributos foram desconsiderados. Após remover estas características, um total de 75 características foi considerado para tarefa de classificação textual (ver Tabela 2).

3.3. Métodos 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 and 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. A parametrização destes classificadores foi a padrão fornecida pelo scikit-learn v1.1, mudando apenas o número de estimadores da floresta aleatória para 2.

A avaliação dos algoritmos foi feita usando métricas já bem conhecidas para classificação [Hossin and Sulaiman 2015]: Precisão, Revocação e F1. Para o entendimento

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

destas métricas é necessário compreender os 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). Desta forma, cada uma das métricas é calculada através das seguintes equações:

$$\textit{Precisão} = \frac{TP}{TP + FP}, \tag{1}$$

$$\textit{Revocação} = \frac{TP}{TP + FN}, \tag{2}$$

$$\textit{Pontuação F1} = \frac{2 * \textit{Precisão} * \textit{Revocação}}{\textit{Precisão} + \textit{Revocação}}. \tag{3}$$

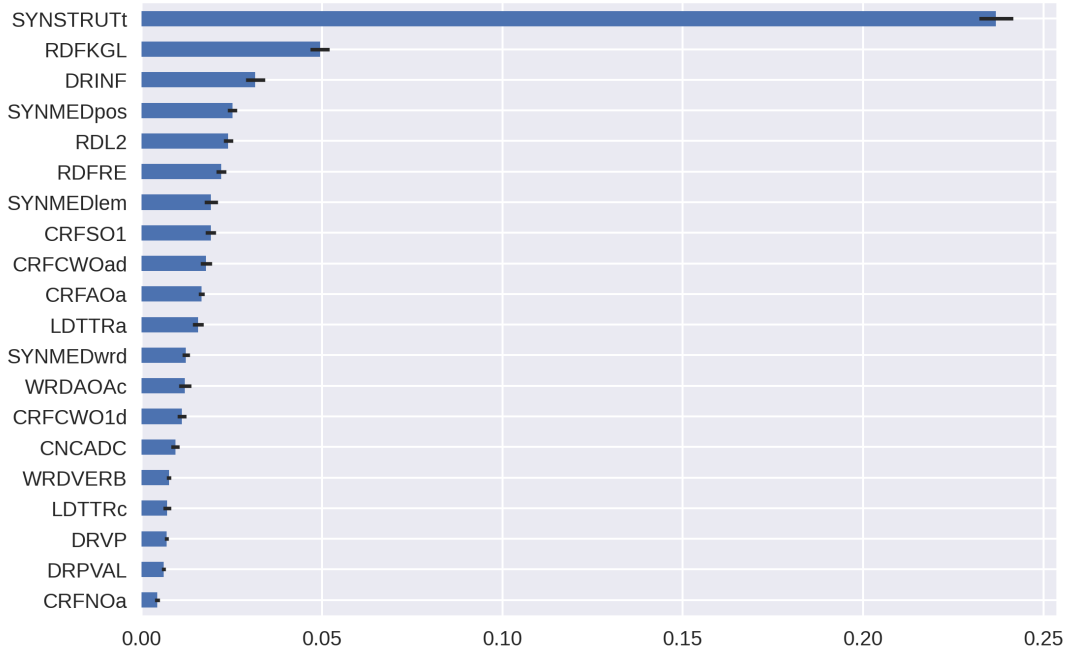
Como se pode ver, a precisão (Equação 1) calcula o percentual de acertos da classe positiva diante de todas as predições positivas. A revocação (ou cobertura) (Equação 2), indica o quanto o modelo está identificando os casos positivos corretamente. Por fim, a pontuação f1 (Equação 3) calcula a média harmônica entre a precisão e a revocação.

4. Resultados e Discussão

4.1. Seleção de Características

Ao todo foram consideradas 75 características no processo de classificação textual. Todas as características realmente contribuem para a classificação? Esta seção detalha o processo de seleção de características adotado. Dentre as mais variadas formas de seleção de características [Chandrashekar and Sahin 2014], utilizamos a seleção baseada na permutação proposta por [Altmann et al. 2010]. Para isso, foi adotado o valor 30 como semente de geração aleatória e o modelo de floresta aleatória para cálculo da importância. Os valores médio e de desvio padrão das importâncias dos índices textuais são demonstrados na Figura 1.

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



Como se pode ver na Figura 1, a característica *SYNSTRUTt* apresentou a maior importância classificação textual. A SYNSTRUTt calcula a semelhança entre as árvores sintáticas de duas sentenças. Assim, SYNSTRUTt obtém a média de todas as combinações de árvores de um texto, ou seja, informa o quanto a estrutura de um texto é uniforme entre si (ver Equação 4).

$$SYNSTRUTt = \frac{nós\ em\ comum}{total\ de\ nós\ -\ nós\ em\ comum} \tag{4}$$

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. Isto pode contribuir para aumentar a diferença entre os textos.

4.2. Análise dos Modelos de Classificação

A Tabela Tabela 3 apresenta os resultados médios obtidos por cada um dos algoritmos em termos de precisão, revocação e F1. Os algoritmos foram avaliados em 3 diferentes condições: 1) usando apenas o SYNSTRUTt como única característica para classificação textual; 2) usando as 20 características de maior importância; e 3) usando todas as 75 características.

Os resultados mostram que os modelos em geral obtiveram bom desempenho na detecção dos tipos textuais. Os modelos treinados utilizando apenas o índice SYNSTRUTt, especificamente, alcançaram resultados superiores aos seus pares quando consideramos as métricas revocação e 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 uma e 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
	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
	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
	F1	**0.9786**	0.0068	0.9037	0.0527	0.9383	0.0419

No geral os três modelos possuem ótimos resultados na classificação dos tipos textuais no TTBR, com métricas de precisão e revocação acima de 95%. O índice SYNSTRUTt [Camelo et al. 2020] como demonstrado na Figura 1, possui a maior relevância na classificação. O experimento usando o modelo de RF mostra também um aumento na precisão no experimento com 20 índices o que mostra uma contribuição das outras características textuais para a predição, este modelos atingiu 99,1% de precisão com um desvio padrão de 0,5%, sendo a melhor métrica obtida no TTBR.

4.3. Limitações

A maior limitação do presente trabalho é o número limitado de exemplos argumentativos no TTBR. Este número precisou ser reduzido para manter o corpus balanceado. Além disso, a falta de textos de outros tipos textuais: expositivo e injuntivo faz com que a área de atuação do classificador fique limitada.

Outra possível limitação é o possível *overfitting* dos modelos por causa da baixa variabilidade nos atributos preditores descritivos de textos narrativos. Ou seja, os atributos como número de linhas e parágrafos têm um peso grande caracterização destes textos, tornando a distinção das classes mais simples. Isso tornou-se ainda mais aparente por conta do pré-processamento realizado a fim de tornar o TTBR balanceado em termos de textos, palavras e caracteres.

5. Conclusões

Este trabalho apresentou um novo corpus de tipos textuais em português, denominado Corpus TTBR, trazendo inicialmente textos dissertativos e narrativos. Além de fazer a

seleção das *features* do CohMetrix com maior importância para o problema de classificação de tipos textuais levando a uma comparação onde o índice SYNSTRUTt demonstra alta capacidade de generalização. E por fim uma análise experimental usando algoritmos *machine learning* clássicos para a classificação dos *tipos textuais* no português, os resultados demonstraram alta performance em todos os casos de teste. Em trabalhos futuros o *dataset* pode ser incrementado com novos exemplos de *tipos textuais* aumentando e encorpando os experimentos, uma vez que já que os modelos obtiveram alto nível de precisão, o questionamento de como fica a performance para um corpus completo, considerando todos os 4 tipos textuais, com estes modelos.

Referências

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Awad, M. and Khanna, R. (2015). Support vector machines for classification. In *Efficient learning machines*, pages 39–66. Springer.

Balint, M., Dascalu, M., and Trausan-Matu, S. (2016). Classifying written texts through rhythmic features. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 121–129. Springer.

Botta, A., de Donato, W., Persico, V., and Pescapé, A. (2016). Integration of cloud computing and internet of things: A survey. *Future Generation Computer Systems*, 56:684–700.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Camelo, R., Justino, S., and de Mello, R. F. L. (2020). Coh-metrix pt-br: Uma api web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, pages 179–186. SBC.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., and Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 404–414.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., and Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, 4(1).

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis.

Ke, Z. and Ng, V. (2019). 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*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4).

Lagutina, K. and Lagutina, N. (2021). A survey of models for constructing text features to classify texts in natural language. In *2021 29th Conference of Open Innovations Association (FRUCT)*, pages 222–233.

Li, S., Xu, L. D., and Zhao, S. (2015). The internet of things: a survey. *Information Systems Frontiers*, 17(2):243–259.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Melissourgou, M. N. and Frantzi, K. T. (2017). Genre identification based on sfl principles: The representation of text types and genres in english language teaching material. *Corpus Pragmatics*, 1(4):373–392.

Mustonen, S. (1965). Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics*, 4:37–44.

Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*, 46(2):330–348.

Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, 44(1):28–47.

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., and Belfkih, S. (2018). Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4):431–448.

Patout, P.-A. and Cordy, M. (2019). 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*, EASEAI 2019, page 17–20, New York, NY, USA. Association for Computing Machinery.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Travaglia, L. C. (2002). Tipos, gêneros e subtipos textuais e o ensino de língua materna. *Língua Portuguesa: uma visão em mosaico. São Paulo: EDUC*, pages 201–214.

Travaglia, L. C. (2003). Tipelementos e a construção de uma teoria tipológica geral de textos. *FÁVERO, Leonor Lopes; BASTOS, Neusa M. de O. Barbosa*, pages 97–117.

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

Wachsmuth, H. and Bujna, K. (2011). Back to the roots of genres: Text classification by language function. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 632–640.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

Zhou, X., Gururajan, R., Li, Y., Venkataraman, R., Tao, X., Bargshady, G., Barua, P. D., and Kondalsamy-Chennakesavan, S. (2020). A survey on text classification and its applications. *Web Intelligence*, 18:205–216. 3.