Detecção Automática de Clímax em Produções de Textos Narrativos

Hyan H. N. Batista[1], Gabriel A. Barbosa[1], Péricles Miranda[1], Jário Santo[5], Seiji Isotani[4,5], Thiago Cordeiro[3], Ig Ibert Bittencourt[3], 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 automatic correction of essays is a problem that has been extensively explored in recent years. One of the most challenging aspects of this task is the assessment of the student's level of mastery of the most varied types of textual structures. Narrative structure is an especially complex case because of its extremely subjective character. Previous works in the area of textual correction did not address the problem of automating the assessment of the student's level of competence in narrative writing. Previous works in the field of textual correction in Portuguese generally analyze essay and not narrative texts. Therefore, there is an important gap in the field. This work investigates the use of machine learning algorithms for climax detection in Portuguese essays as an initial step in solving the problem of automatic correction of narrative texts. Three traditional classification algorithms, the support vector machine, random forest and stochastic gradient descent, were applied to an annotated dataset translated into Portuguese. The algorithms were evaluated in terms of accuracy, recall and F_1 score, with random forest being the best performing algorithm. In addition, an analysis of the attributes involved was performed, and the experiments showed that the best results are obtained when combining attributes from both the Coh-Metrix and the LIWC.*

Resumo. *A correção automática de redações é um problema que vem sendo bastante explorado nos últimos anos. Um dos aspectos mais desafiadores nessa tarefa é a avaliação do nível de domínio do aluno quanto aos mais variados tipos de estruturas textuais. A estrutura narrativa é um caso especialmente complexo devido ao seu caráter extremamente subjetivo. Trabalhos anteriores na área de correção textual em português, em geral analisam textos dissertativos e não narrativos. Portanto, existe uma lacuna importante na área. Este trabalho*

DOI: 10.5753/sbie.2022.224770

investiga o uso de algoritmos de aprendizagem de máquina para a detecção de clímax em redações em português como um passo inicial na resolução do problema de correção automática de textos narrativos. Três algoritmos de classificação tradicionais, o support vector machine, floresta aleatória e descida de gradiente estocástica, foram aplicados em um conjunto de dados anotado traduzido para o Português. Os algoritmos foram avaliados em termos de precisão, revocação e pontuação F_1, sendo a floresta aleatória o algoritmo de melhor desempenho. Além disso, foi realizada uma análise dos atributos envolvidos, e os experimentos mostraram que os melhores resultados são obtidos ao combinar-se atributos tanto do Coh-Metrix quanto do LIWC.

1. Introdução

A correção de produções textuais no contexto educacional é algo que exige do corretor alta expertise, tempo e concentração. Dependendo da quantidade de textos e de profissionais envolvidos, ela pode demorar desde dias até semanas ou meses. Entretanto, existe uma tendência que está crescendo exponencialmente em se aplicar métodos computacionais e estatísticos para automatizar esse processo [Ramesh and Sanampudi 2021, Passero et al. 2019].

Por exemplo [Wang et al. 2018] propôs um sistema automático de pontuação (AES, do inglês *automatic essay scoring*) de redações baseando-se em técnicas de processamento de linguagem natural e aprendizado profundo para ajudar professores. Por sua vez [Fonseca et al. 2018] usou redes neurais profundas para criar um AES que fosse capaz de gerar automaticamente avaliações a respeito das cinco competências requisitadas na redação do ENEM (Exame Nacional do Ensino Médio). Um outro estudo, dessa vez voltado para textos narrativos, feito por [Ouyang and McKeown 2014], baseou-se na teoria da análise de narratividade proposta por [Labov and Waletzky 1967] e realizaram análises pioneiras na detecção de estruturas narrativas.

Apesar dos trabalhos anteriores na correção de textos dissertativos [Costa et al. 2020], a avaliação automática de textos narrativos ainda é um tópico complexo e pouco analisado, tanto pela escassez de dados, quanto pela natureza subjetiva da tarefa. Teorias como a de [Labov and Waletzky 1997] e o estudo realizado por [Ouyang and McKeown 2014] dão uma direção inicial rumo à solução do problema. Contudo, eles não abordam o problema de automatizar a avaliação do nível de domínio que o estudante possui na construção de narrações, nem a detecção de clímax. Além disso, os experimentos foram feitos em narrativas orais em inglês de experiências pessoais. Portanto, os modelos gerados não se aplicam a estudantes nativos de países de língua portuguesa.

Diante deste cenário, este trabalho propõe um modelo que pode automaticamente analisar textos narrativos em português. Mais precisamente, este artigo introduz um conjunto de classificadores binários textuais capazes de detectar a presença de clímax. O método proposto atingiu 94% e 95% de acurácia e F_1, respectivamente. Os modelos produzidos neste estudo podem ser utilizados na construção de ferramentas de correção automática de narrações para facilitar o trabalho de professores de escolas públicas e particulares.

2. Trabalhos relacionados

Ao longo das últimas duas décadas houve um desenvolvimento significativo na área de avaliação automática de textos. Algoritmos de aprendizado de máquina e *deep learning* foram empregados na implementação de diversos sistemas de correção automática de textos com os mais variados propósitos [Vijaya Shetty et al. 2022].

Um dos primeiros sistemas de pontuação automática de redações foi o PEG (do inglês, Project Essay Grader) [Rudner and Gagne 2000]. Esse sistema focava em avaliar a qualidade de escrita do autor utilizando algoritmos de regressão linear múltipla e era capaz de gerar pontuações para textos não previamente avaliados. Entretanto, a sua principal limitação é que o método empregado não utiliza características contextuais [Vijaya Shetty et al. 2022].

Remediando essa limitação específica do PEG, o IEA (do inglês, Intelligent Essay Assessor), um outro AES (*Automated Essay Scoring*) forneceu uma solução para o problema de pontuação da qualidade do conteúdo do texto escrito [Foltz et al. 1999]. Para esse propósito esse sistema empregou técnicas de análise de semântica latente. Dessa forma, ele consegue capturar relações de transitividade e quantificar seu conteúdo semântico. Um ponto negativo nesse AES é que ele não disponibiliza devolutivas para que o autor aprimore a sua escrita [Vijaya Shetty et al. 2022].

Por outro lado, o *E-rater* usa técnicas de processamento de linguagem natural e estatística para mensurar habilidade de escrita junto com questões dissertativas que requerem respostas curtas [Burstein et al. 2001]. Dessa maneira ele consegue fornecer devolutivas a respeito da gramática, mecanismos de escrita, estilo e organização, apresentando uma acurácia de 87%. A principal limitação desse sistema é que ele assume que não há diferença entre textos bons e ruins [Vijaya Shetty et al. 2022].

Por sua vez, [Wang et al. 2018] combinaram as estruturas de uma rede neural BiLSTM e uma rede de atenção hierárquica para treinar um modelo usando os dados do conjunto *Automated Student Assessment Prize* do *Kaggle* a fim de criar um sistema fim-a-fim capaz de gerar pontuações de forma automática com o intuito de auxiliar o professores na tarefa de correção das redações. O sistema desenvolvido obteve desempenho de estado-da-arte, atingindo 83% de concordância com os avaliadores humanos. Além disso, o emprego de mecanismos de atenção deu ao modelo a habilidade de focar nas partes ilógicas da redação e julgar a relação lógica correta entre cada uma das sentenças.

No que se refere à textos narrativos, [Ouyang and McKeown 2014] usou características narrativas, de discurso e compartilhadas como input para um algoritmo de regressão logística. Apesar do trabalho ser relacionado, ele foi realizado com narrações em inglês.

O presente trabalho apresentado neste artigo baseia-se em pesquisas anteriores para explorar o uso de métodos de processamento de linguagem natural baseados em aprendizado de máquina como solução para o problema de detecção automática de clímax em redações narrativas. Enquanto os trabalhos anteriores focaram no uso de técnicas computacionais diversas para gerar pontuações para redações, este trabalho foca em textos narrativos, mais especificamente, em um de seus principais componentes, o clímax. Os resultados desse trabalho poderão ser usados na construção de sistemas computacionais capazes de realizar automaticamente a correção de redações narrativas.

3. Perguntas de pesquisa

Como dito nas seções anteriores, é crucial o desenvolvimento de métodos de detecção automática de clímax que auxiliem os examinadores na tarefa de corrigir textos narrativos. Embora tenham sido feitos vários estudos na área de análise narrativa, a literatura não dispõe de uma gama ampla de trabalhos que realizam a detecção automática de eventos altamente reportáveis. Por essa razão, este artigo propõe um estudo dos preditores extraídos por essas ferramentas, o que nos leva à nossa primeira pergunta de pesquisa:

Pergunta de Pesquisa 1:

Até que ponto métodos de aprendizado de máquina conseguem identificar precisamente e automaticamente a presença de clímax em um texto narrativo?

Para além da detecção automática da presença de clímax, também é intenção deste trabalho fornecer informações adicionais sobre os preditores mais importantes para a classificação de narrações no que se refere à presença de eventos altamente reportáveis. Para tanto, explorou-se o uso do teste de independência Chi^2 (*do inglês, Chi-square test of independence*).

Pergunta de Pesquisa 2:

Quais os melhores preditores para detectar-se a presença de clímax em narrações?

4. Método

4.1. Dados

O *data set* utilizado para executar os experimentos apresentados neste artigo foi baseado no conjunto de dados construído por [You and Goldwasser 2020] é chamado de *Social Tree Narrative* (STN). O STN contém textos em inglês dividido em cinco componentes, ou classes. São eles a semente, a construção, o clímax, a resolução e o desfecho. Quando juntos, esses componentes formam o que os autores denominaram de narrativa social. O *data set* é constituído de dez sementes, 5 construções, 5 clímax e 5 pares resolução-desfecho que foram combinados, totalizando 1250 narrativas sociais completas. O conjunto STN foi criado para que os autores pudessem treinar modelos computacionais capazes de predizer o desfecho de uma conversa com base nos outros componentes já citados.

Para cada uma das narrativas sociais presentes nesse conjunto de dados existem anotações a respeito dos seus eventos mais reportáveis, assim como definido por [Labov and Waletzky 1967]. Essa característica o torna adequado para a tarefa de detecção automática de clímax em textos narrativos. Contudo, como os textos não estavam em português, então, eles foram traduzidos para o português brasileiro usando o Google Translator [1]. Para cada um dos textos gerados nesse processo, foram criadas uma amostra positiva e uma negativa de redação que continha ou não o trecho relacionado ao clímax. A Tabela 1 mostra a distribuição das classes dos dados após a execução das operações descritas anteriormente.

[1] https://translate.google.com/

Table 1. Distribuição das classes do conjunto de dados após as transformações.

Classe 0	Classe 1	Total
1250	1250	2500
1250	1250	2500

4.2. Extração das características

Na área de mineração de textos educacionais, a maioria dos trabalhos iniciais empregavam o uso de características léxicas *N-gram* ou similares, como bigramas de classes gramaticais ou triplas de dependência [Ferreira-Mello et al. 2019]. Contudo, trabalhos recentes [Kovanović et al. 2016, Cavalcanti et al. 2020] mostraram que: (i) características desse tipo inflam o espaço de características, até mesmo para pequenos conjuntos de dados, o que aumenta em muito as chances de acontecer um *over-fitting* nos dados de treino [Kovanović et al. 2016]; (ii) essas características são são altamente dependentes do *data set*, visto que os dados em si definem o espaço de classificação [Kovanović et al. 2016], o que torna muito difícil um conjunto fixo de características de classificação com antecedência, haja que a escola particular de palavras nos documentos de treino vão definir quais características são usados para classificação [Cavalcanti et al. 2020].

Baseado nos resultados estudos anteriores esse trabalho vai avaliar a eficácia da utilização de características do LIWC [Francis and Booth 1993] e do Coh-Metrix [McNamara et al. 2014], assim como a combinação dos características extraídos por ambas as ferramentas. Além disso, posteriormente à etapa extração, foi executada uma etapa de seleção de atributos características com o propósito de se evitar *overfitting* como alertado por [Kovanović et al. 2016]. Apenas os 20 características mais importantes de cada um dos *data sets* foram incorporados em suas versões finais. Nas seções seguintes serão apresentados todos os características selecionadas para este estudo.

4.2.1. LIWC

O LIWC (do inglês, Liguistic Inquiry Word Count) é um dicionário léxico [Francis and Booth 1993]. Ele agrupa palavras em categorias com significado psicológico, como emoções, processos cognitivos, preocupações pessoais e palavras sociais, mas também em categorias que consideram aspectos gramaticais e linguísticos como verbos, preposições e conjugações em primeira, segunda e terceira pessoas. O seu maior diferencial, entretanto, está em sua capacidade de fornecer informações a respeito do estado psicológico ou condição pessoal do autor [Van Wissen and Boot 2017]. De acordo com [Balage Filho et al. 2013] e [Kovanović et al. 2016], o LIWC tem 127.149 entradas, onde cada entrada pode ser atribuída a uma ou mais categorias e, para além disso, fornece um grande número de contagens de palavras que são indicativas de diferentes processos psicológicos, tais como afetivo, cognitivo, social e perceptivo. Seguindo a metodologia de [Cavalcanti et al. 2020], para este trabalho nós extraímos 64 características. Como é no clímax que a narração atinge o seu pico de intensidade emocional, características que descrevem processos cognitivos, psicológicos e emocionais se tornam bastante relevantes para o problema aqui tratado, fornecendo descrições para o texto que vão além de seu conteúdo.

Table 2. Distribuição das classes dos *datasets* após a divisão.

	Classe 0	Classe 1	Total
Treino	938 (50,03%)	937 (49,96%)	1875
Teste	313 (50,08%)	312 (49.92%)	625

4.2.2. Coh-Metrix

O Coh-Metrix é uma ferramenta linguística computacional que calcula índices que avaliam a coesão, coerência e dificuldade de compreensão de um texto usando vários níveis de análise linguística, tais como léxico, sintático, discursivo e conceitual [Scarton and Aluísio 2010]. O *Coh-Metrix PT-BR* utiliza modelos estatísticos e computacionais suportados por ferramentas como o *spaCy* para extrair dados linguísticos de textos em português brasileiro para propósitos educacionais [Camelo et al. 2020]. Essas medidas permitem uma análise profunda do conteúdo do texto do ponto vista linguístico. Como o propósito deste trabalho é apresentar um método de classificação de textos com clímax, o Coh-Metrix torna-se uma ferramenta ainda mais atrativa, visto que fornece um conjunto de características extraídos do texto que são amplamente adotados na literatura educacional para avaliar a qualidade dos textos e atividades escritas [McNamara et al. 2014].

4.3. Processamento dos dados

Seguindo os princípios estabelecidos por [Hastie et al. 2009], para que se evitasse uma estimação superestimada da performance do modelo, inicialmente, os *datasets LIWC*, *Coh-Metrix* e *Coh-Metrix + LIWC* foram divididos em conjuntos de treino e teste, seguindo uma proporção de 75% e 25% respectivamente. Esse procedimento foi realizado usando um algoritmo de estratificação que mantém a proporção de exemplos positivos e negativos o mais próximo possível para cada um dos conjuntos gerados através de uma seleção aleatória das amostras. Isto previne *overfitting* e aumenta a capacidade de generalização dos modelos produzidos [Hastie et al. 2009]. Dessa forma, os dados gerados dessa estratificação continham 1875 amostras para o conjunto de treino e 625 para o conjunto de teste para cada um dos *datasets* mencionados. Na Tabela 2, é pode-se observar a divisão para cada um dos conjuntos de dados. Como é possível observar, os dados estão balanceados, haja vista que os exemplos negativos foram criados a partir da remoção do clímax das amostras positivas. Por essa razão, a utilização de técnicas de balanceamento de conjuntos de dados não foi necessária.

4.4. Seleção de modelo e avaliação

Para classificar automaticamente os textos em contendo clímax foram utilizados diferentes algoritmos. O gradiente descendente estocástico tem sido amplamente aplicado em diversos problemas de aprendizado de máquina, como classificação e predição linear, com resultados satisfatórios [Rajkumar and Agarwal 2012, Al-Anzi 2022b, Al-Anzi 2022a]. De forma similar ao perceptron, ele atualiza o vetor de pesos w conforme é alimentado com amostras de treino, em outras palavras, é capaz de performar aprendizado de máquina *online* [Zhang 2004]. Por outro lado, florestas aleatórias e máquinas de vetores de suporte, como demonstrado por [Fernández-Delgado et al. 2014], em uma análise compar-

ativa contando com 179 algoritmos e 121 conjuntos de dados distintos, apresentam os melhores resultados em tarefas de classificação. Por essa razão, neste estudo, para abordarmos a primeira pergunta de pesquisa foram realizados experimentos a fim de identificar qual algoritmo melhor modelaria os dados.

Para avaliar o desempenho desses algoritmos utilizamos as seguintes métricas: precisão, cobertura, F_1, acurácia e *Cohen's Kappa*. A precisão é usada para mensurar as amostras positivas que foram corretamente classificadas do total de classificações em uma classe positiva. A cobertura, por outro lado, mede a fração de amostras positivas que foram corretamente classificadas. A média harmônica dessas duas métricas é chamada de F_1 [Hossin and Sulaiman 2015]. A acurácia mensura a razão entre as predições corretas, sejam elas positivas ou negativas, e o número total de instâncias avaliadas [Hossin and Sulaiman 2015]. Finalmente, o *Cohen's Kappa* é uma métrica utilizada para analisar o nível de concordância entre dois anotadores que, neste contexto, seriam o classificador e as anotações manuais [Cohen 1960].

Por fim, foi utilizada a técnica de teste de independência Chi [McHugh 2013] para avaliar a importância das características analisadas para o problema de identificação de clímax. Essa técnica é capaz de relacionar a um atributo uma pontuação que diz respeito ao seu grau de relevância ou, em outras palavras, o quão influente o valor daquele atributo foi no processo de classificação.

4.5. Implementação

As extração dos preditores e classificação foram feitas usando a linguagem Python. Os pacotes e bibliotecas utilizados foram:

- *scikit-learn*[2], para criação da amostragem estratificada dos conjuntos de treino e teste, como também, para o treinamento dos modelos de classificação;
- *Coh-Metrix PT-BR* [Camelo et al. 2020] e;
- A versão para língua portuguesa do *LIWC*[3]

5. Resultados e discussão

5.1. Avaliação dos modelos de predição - PP1

Para o primeiro experimento foi analisado cada conjunto de características apresentados nos classificadores descritos nas seções anteriores. A Tabela 3 mostra que para todas as configurações de características o modelo que obteve o melhor desempenho considerando todas as métricas citadas anteriormente, foi a floresta aleatória, seguindo da máquina de vetores de suporte e, finalmente, o gradiente de descida estocástico. Esses resultados corroboram os achados de [Fernández-Delgado et al. 2014], reforçando ainda mais a superioridade dos algoritmos floresta aleatória e máquina de vetores de suporte em tarefas de classificação quando comparados com os demais.

Para mais, os resultados apontam, também, uma notória diferença de performance para cada um dos algoritmos dependendo do conjunto de características no qual foi treinado. Os melhores resultados foram obtidos usando-se os atributos de ambos, do LIWC e do Coh-Metrix. Contudo, há uma diferença quando analisados isoladamente. Os

[2]https://scikit-learn.org/

[3]http://www.nilc.icmc.usp.br/portlex/index.php/en/liwc

atributos do LIWC são em geral menos descritivos para identificação de clímax e, portanto, preditores mais não tão eficientes quando comparados com os do Coh-Metrix. O gradiente descendente estocástico treinado com os índices do Coh-Metrix, por exemplo, tem o pior desempenho entre os três algoritmos para esse conjunto. Entretanto, obteve pontuações superiores ao modelo com melhor desempenho do LIWC, a floresta aleatória.

	LIWC			**Coh-Metrix**			**Coh-Metrix + LIWC**		
	SVM	**SGD**	**RF**	**SVM**	**SGD**	**RF**	**SVM**	**SGD**	**RF**
Precisão	0.838	0.807	0.858	0.938	0.919	0.946	0.946	0.935	**0.951**
Cobertura	0.837	0.807	0.858	0.938	0.917	0.946	0.946	0.935	**0.949**
F_1	0.837	0.808	0.858	0.938	0.917	0.946	0.946	0.936	**0.949**
Acurácia	0.842	0.802	0.845	0.933	0.936	**0.946**	**0.946**	0.933	0.941
Cohen's kappa	0.701	0.701	0.717	0.877	0.837	0.890	0.893	0.890	**0.901**

Table 3. Pontuações dos modelos para cada um dos modelos e conjuntos de dados.

5.2. Análise de importância das características - PP2

Além de realizar análises comparativas de algoritmos de aprendizado de máquina dentro do contexto de detecção de clímax, este estudo também analisou a importância que as diferentes características extraídas utilizando o LIWC e o Coh-Metrix têm na performance dos modelos preditivos gerados.

Neste trabalho analisamos as características extraídas do LIWC e do Coh-Metrix simultaneamente. Essa combinação dos aspectos linguísticos identificados por cada ferramenta permite que o grau de importância delas seja analisado em um contexto onde todas estão juntas e então explorar os benefícios que isso pode gerar em questão de performance para os modelos construídos. Observando os dados da Tabela 5.2, pode-se notar que o índice mais relevante nas classificações é o *liwc.we*, o número de vezes que o pronome da primeira pessoa do plural foi utilizado. Contudo, logo em seguida, com diferença muito pequena, vêm os índices de coesão referencial, o *coh.CRFAOa* e *coh.CRFSOa*, sobreposição de argumentos e raízes, respectivamente. É possível supor, então, a partir dessa análise, que para a tarefa de detecção de clímax, dada a disponibilidade de índices provenientes de ambas as ferramentas, índices que deem informações sobre as dimensões linguísticas e coesão referencial do texto a ser processado são os que apresentam maior grau de importância. Por fim, é importante destacar que do top-20 características 13 são do Coh-Metrix e 7 do LIWC.

6. Discussão

A PP1 tinha o objetivo de investigar até que ponto métodos de aprendizado de máquina e processamento de linguagem natural conseguem classificar automaticamente textos narrativos quanto à presença de clímax. Os resultados mostraram que os classificadores treinados com os atributos preditores extraídos com o LIWC e o Coh-Metrix foram efetivos na identificação do clímax. Eles, também, mostraram que o uso dessas duas ferramentas produz modelos mais performáticos que quando usados separadamente, atingindo 90,1% de concordância com as anotações manuais. Quando analisados isoladamente, contudo,

#	Variável	Descrição	Pontuação
1	liwc.we	Incidência de primeira pessoa do plural	78,381
2	coh.CRFAOa	Sobreposição de argumentos	68,178
3	coh.CRFSOa	Sobreposição de lemas em todas as sentenças	59,134
4	coh.DESPL	Tamanho médio dos parágrafos	57,852
5	coh.DESSC	Número de sentenças	57,852
6	coh.CRFSO1	Sobreposição de lemas em sentenças adjacentes	50,093
7	coh.CRFNOa	Sobreposição de substantivos em todas as sentenças	48,971
8	coh.DESWC	Número de palavras	47,623
9	coh.CRFAO1	Sobreposição de argumentos em sentenças subjacentes	47,531
10	liwc.funct	Incidência de palavras de função	46,571
11	liwc.incl	Incidência de palavras de inclusão	39,449
12	liwc.cogmech	Incidência de palavras categorizadas em processos cognitivos	37,549
13	liwc.past	Incidência de palavras no tempo verbal passado	37,463
14	coh.SMINTEp_sentence	Incidência de verbos intencionais	37,442
15	coh.CRFNO1	Sobreposição de substantivos em sentenças subjacentes	37,272
16	liwc.verb	Incidência de verbos	35,813
17	liwc.relativ	Incidência de palavras de relatividade	35,612
18	coh.WRDVERB	Incidência de verbos	35,586
19	coh.DRVP	Incidência de verbos frasais	35,300
20	coh.WRDNOUN	Incidência de substantivos	34,614

Table 4. Os 20 características mais importantes para o conjunto de dados Coh-Metrix + LIWC

os atributos extraídos pelo Coh-Metrix geraram modelos melhores que os do LIWC, apresentando uma diferença de 17,3%, com relação ao *Cohen's Kappa*, quando comparado com os melhores classificadores para cada uma delas.

O foco da PP2 foi selecionar os atributos mais significados para a detecção de clímax em redações narrativas. Para respondê-la, aplicou-se o teste de independência Chi na construção de tabelas que contivessem os 20 atributos mais importantes. Os dados mostraram que, para o LIWC, as incidências de artigos, pronomes, preposições, verbos auxiliares, advérbios comuns, conjunções e primeira pessoa do plural são os atributos mais importantes. Para o Coh-Metrix, por outro lado, o número de sobreposições de lemas e argumentos, seguidos de atributos que descrevem o texto estruturalmente, como tamanho médio dos parágrafos e número de sentenças, são os mais significativos. Quando usados juntos, entretanto, atributos referentes à coesão referencial, estrutura do texto e suas dimensões linguísticas são as mais relevantes.

7. Conclusão

Este estudo produziu duas contribuições principais para a literatura na área. A primeira foi um estudo comparativo do desempenho de diferentes modelos de aprendizado de máquina como solução para o problema de detecção da presença de clímax em textos narrativos. Esses modelos foram treinados em textos traduzidos do inglês para o português retirados da base de dados produzida por [You and Goldwasser 2020]. Os resultados corroboraram com estudos anteriores [Fernández-Delgado et al. 2014] mostrando que florestas aleatórias e máquinas de vetores de suporte são os melhores algoritmos não baseados em redes neurais para tarefas de classificação. Além disso, o método proposto, que se baseia

na extração de atributos preditores usando o Coh-Metrix e o LIWC em suas versões para português brasileiro demonstrou que é possível desenvolver um sistema automático de detecção da presença de clímax em narrações.

A segunda contribuição foi a realização de uma análise dos atributos mais importantes, fornecendo um entendimento mais aprofundado sobre quais são os preditores mais fortes de presença de clímax em narrações. Essa análise foi fundamental para demonstrar a superioridade das métricas do Coh-Metrix quando comparadas com as do LIWC para a tarefa de classificação textual proposta, mesmo que, quando juntas, apresentam pontuações melhores do que quando usadas isoladamente.

Algumas limitações do estudo são importantes de serem apresentadas. A primeira delas é o fato de que os textos usados foram traduzidos do inglês para o português. Por essa razão, eles podem conter frases gramaticalmente incorretas ou até com expressões que foram traduzidas literalmente e que, portanto, não fazem sentido contextualmente. Isso pode diminuir a capacidade de generalização do modelo. Uma outra limitação do estudo, é sobre a variabilidade de sentenças de clímax. Isso pode ser problemas, pois se elas são pouco variadas, a estratificação do conjunto de dados se torna um método menos eficiente para fins de validação e, como as amostras negativas foram geradas por meio da remoção de uma sentença do texto: o clímax, o número de parágrafos se torna um atributo não confiável, porém, significativo nas predições.

Como trabalhos futuros, pretende-se alocar esforços na construção de um conjunto de dados de narrações em português brasileiro que possibilite o uso de técnicas de processamento de linguagem natural, como rotulação de sequências e classificação textual, não só para a detecção de clímax, mas também de outros componentes típicos de um texto narrativo. Além disto, visamos criar uma aplicação que possa ser usada por professores e alunos seguindo princípios de Learning ANalytics [Sousa et al. 2021, Freitas et al. 2020].

References

Al-Anzi, F. S. (2022a). An effective hybrid stochastic gradient descent arabic sentiment analysis with partial-order microwords and piecewise differentiation. In *2022 9th International Conference on Electrical and Electronics Engineering (ICEEE)*, pages 408–411. IEEE.

Al-Anzi, F. S. (2022b). An effective hybrid stochastic gradient descent for classification of short text communication in e-learning environments. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, volume 1, pages 1096–1101. IEEE.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Burstein, J., Leacock, C., and Swartz, R. (2001). Automated evaluation of essays and short answers.

Camelo, R., Justino, S., and Mello, R. (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, Porto Alegre, RS, Brasil. SBC.

Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., and Gašević, D. (2020). How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 428–437.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costa, L., Oliveira, E., and Júnior, A. C. (2020). Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412, Porto Alegre, RS, Brasil. SBC.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

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.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *International Conference on Computational Processing of the Portuguese Language*, pages 170–179. Springer.

Francis, M. and Booth, R. J. (1993). Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Freitas, E., Falcão, T. P., and Mello, R. F. (2020). Desmistificando a adoção de learning analytics: um guia conciso sobre ferramentas e instrumentos. In *Jornada de Atualização em Informática na Educação*, pages 73—-99. Sociedade Brasileira de Computação.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

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.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., and Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 15–24.

Labov, W. and Waletzky, J. (1967). Narrative analysis. essays on the verbal and visual arts, ed. by june helm, 12-44.

Labov, W. and Waletzky, J. (1997). Narrative analysis: Oral versions of personal experience.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

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.

Ouyang, J. and McKeown, K. (2014). Towards automatic detection of narrative structure. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4624–4631, Reykjavik, Iceland. European Language Resources Association (ELRA).

Passero, G., Ferreira, R., and Dazzi, R. L. S. (2019). Off-topic essay detection: A comparative study on the portuguese language. *Revista Brasileira de Informática na Educação*, 27(03):177–190.

Rajkumar, A. and Agarwal, S. (2012). A differentially private stochastic gradient descent algorithm for multiparty classification. In *Artificial Intelligence and Statistics*, pages 933–941. PMLR.

Ramesh, D. and Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, pages 1–33.

Rudner, L. M. and Gagne, P. (2000). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation*, 7(1):26.

Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Sousa, E. B. d., Alexandre, B., Ferreira Mello, R., Pontual Falcão, T., Vesin, B., and Gašević, D. (2021). Applications of learning analytics in high schools: a systematic literature review. *Frontiers in Artificial Intelligence*, 4:737891.

Van Wissen, L. and Boot, P. (2017). An electronic translation of the liwc dictionary into dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 703–715. Lexical Computing.

Vijaya Shetty, S., Guruvyas, K., Patil, P. P., and Acharya, J. J. (2022). Essay scoring systems using ai and feature extraction: A review. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*, pages 45–57. Springer.

Wang, Z., Liu, J., and Dong, R. (2018). Intelligent auto-grading system. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 430–435. IEEE.

You, K. and Goldwasser, D. (2020). " where is this relationship going?": Understanding relationship trajectories in narrative text. *arXiv preprint arXiv:2010.15313*.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116.