Classificador de legibilidade de textos em língua inglesa

Levi M. G. Sange[1, *]

[1] Departamento Acadêmico de Computação – Universidade Tecnológica Federal do Paraná (UTFPR) – Medianeira – PR – Brasil

*levisange@alunos.utfpr.edu.br

Abstract. *People interested in learning the English language look for free auxiliary tools, however, despite the large number of existing tools for this purpose, learners still have difficulty finding texts and readings in line with their knowledge of the English language. This work aims, through the developed prototype, to complement existing tools, as well as to serve as a basis for the construction of recommendation systems. For this, Natural Language Processing techniques, machine learning algorithms and free datasets were used to create a prototype for classifying the readability of articles present in the Wikipedia and Simple Wikipedia datasets.*

Resumo. *Pessoas com interesse no aprendizado da língua inglesa buscam por ferramentas gratuitas auxiliares, porém, apesar da grande quantidade de ferramentas existentes com este objetivo, aprendizes ainda possuem dificuldade em encontrar textos e leituras niveladas a seu conhecimento na língua inglesa. Esse trabalho visa, através do protótipo desenvolvido, complementar ferramentas existentes, assim como servir de base para a construção de sistemas de recomendação. Para isso foram utilizadas técnicas de Processamento de Língua Natural, algoritmos de aprendizado de máquina e datasets gratuitos na criação de um protótipo classificador de legibilidade de artigos presentes nos datasets Wikipedia e SimpleWikipedia.*

1. Introdução

Proveniente do ocidente germânico, a língua inglesa possui um vocabulário extenso formada a partir da incorporação de outras línguas. As inúmeras transformações e eventos ocorridos mundialmente impulsionaram o inglês a condição de língua globalizada. Segundo [Baker 2018], alguns dos fatores que também contribuíram para a língua inglesa ser utilizada globalmente são sua propagação geográfica e diversidade cultural dos que a falam. Como segunda língua aprendida por uma pessoa permite a comunicação quase mundialmente. Reconhecendo a importância da língua inglesa, o mercado de trabalho tem, muitas vezes, adicionado o conhecimento dela como requisito para oportunidades de emprego. Acredita-se que cientes desta exigência por parte do mercado de trabalho as pessoas tendem a procurar mais e diferentes formas de obter conhecimento e proficiência na língua inglesa. Porém, conforme [British Council 2014], os especialistas, professores e até governo reconhecem que o ensino de inglês na educação básica, privada ou pública, no Brasil, têm formado poucos estudantes com proficiência no idioma, portanto o ensino tem sido resumido a noções iniciais das regras gramaticais, leitura de textos curtos e desenvolvimento da habilidade de testes de múltipla escolha voltados para exames avaliativos. Tendo em vista este cenário em relação ao aprendizado da língua inglesa,

seria de grande benefício o desenvolvimento de um protótipo ou ferramenta capaz de complementar processos já existentes de aprendizagem. Nesse sentido, algoritmo de classificação é uma categoria de aprendizado de máquina supervisionado que possui potencial para, por exemplo, rotular uma leitura ou texto em níveis de legibilidade com base em características presentes em textos, que podem ser obtidas através de bibliotecas e fórmulas. Estas auxiliam os algoritmos de aprendizado na tarefa de classificar a dificuldade de leitura e compreensão de textos. Com o desenvolvimento de um protótipo de classificação de textos é possível expandir o número de ferramentas gratuitas e de fácil acesso para a área de Processamento de Língua Natural e de aprendizagem da língua inglesa. sendo mais um passo positivo para a educação e pesquisa. Esse artigo propõe o desenvolvimento de um protótipo capaz de classificar textos em níveis de dificuldade de leitura. Assim auxiliando usuários no processo de encontrar textos nivelados utilizando técnicas de Processamento de Língua Natural e algoritmos de classificação de aprendizado de máquina treinados sobre *datasets* gratuitos.

O trabalho apresentado está estruturado da seguinte forma: na Seção 2 são apresentados a teoria em que se baseia o desenvolvimento deste trabalho. Na Seção 3 de metodologia são descritos os materiais e métodos utilizados para a realização dos objetivos propostos pelo trabalho. Na Seção 4 são apresentados, analisados e discutidos os resultados obtidos. As conclusões do trabalho serão apresentadas na Seção 5.

2. Fundamentos

Nesta Seção serão descritos o papel e importância da língua inglesa, técnicas de Processamento de Língua Natural úteis no contexto da língua inglesa e objetivo deste trabalho, algoritmos de aprendizado de máquina para entendimento de como é possível para os computadores, aprender e representar conhecimento, gerar resultados, tomada de decisões, técnicas e abordagens diferentes.

2.1. Língua Inglesa

Uma língua atinge nível global quando tem seu papel específico reconhecido em diversos países ao redor do mundo [Melitz 2016]. Neste sentido, a língua inglesa destaca-se devido a não ser falado apenas no país de origem, mas sim como uma segunda língua que foi aos poucos sendo falada em diferentes países ao redor do mundo, ganhando seu lugar especial nas comunidades globalmente. Setores do governo, mídia e redes sociais a utilizam como intermediária com foco de atingir um público maior devido a aceitação, além de muitos países escolherem aplicá-la na educação como uma língua estrangeira. A língua estrangeira mais falada mundialmente, em cerca de 100 países como China, Rússia, Alemanha, Espanha, Egito e Brasil [Melitz 2016].

2.2. Processamento de Língua Natural

Conforme [Covington 2013] Processamento de Língua Natural (PLN) é o uso de computadores para entender as línguas humanas (naturais), como o inglês. É uma área da Inteligência Artificial que estuda como representar, reconhecer, extrair conhecimento e compreender linguagem natural de forma automática pelas máquinas utilizando técnicas, algoritmos e soluções. [Russel and Norvig 2013] afirmam que apesar de outros animais terem mostrado vocabulários com centenas de sinais, apenas os humanos podem se comunicar de maneira confiável em um número ilimitado de diferentes mensagens. Com base na definição exposta, pode-se observar que a tarefa do PLN não é tão fácil para a

máquina devido a sua complexidade, sua ambiguidade e diversos outros fatores que dificultam o processamento de uma língua natural. Entretanto, é uma tarefa viável e que é aprimorada a cada ano com novas tecnologias e métodos.

2.3. Algoritmo *Naive Bayes*

Segundo [Mitchell 1997] o classificador *Naive Bayes* é aplicado a tarefas em que cada instância é descrita por uma conjunção de valores de atributo, onde a função objetivo pode assumir qualquer valor de algum conjunto finito *c*. A diferença entre este método e os demais é que ele é formado sem pesquisa, pois baseia-se na contagem e frequência das combinações de dados de treinamento. O algoritmo Naive Bayes é baseado no teorema de Bayes, dado pela Equação 1. O teorema diz que a probabilidade do evento A ocorrer dado que o evento B ocorreu é calculada da seguinte forma. Primeiro, é obtido o produto entre a probabilidade do evento B ocorrer dado que A ocorreu e a probabilidade do evento A ocorrer. Na sequência, esse produto é dividido pela probabilidade de B.

$$p(A|B) = \frac{p(\mathrm{B}|\mathrm{A})p(A)}{p(B)} \quad (1)$$

2.4. Máquina de Vetor Suporte

Segundo [Goodfellow et al. 2016] as *Support Vector Machines* (SVM) foram inicialmente criadas para resolver problemas linearmente separáveis de classificação binária utilizando da função $wx + b$, predizendo que a classe é positiva caso o resultado desta função seja positivo, assim também a classe é negativa caso o resultado seja negativo. O algoritmo busca determinar o hiperplano separador ótimo entre instâncias de ambas as classes. Para isso, são utilizadas instâncias de apoio, próximas a fronteira de decisão. A busca pelo hiperplano separador ótimo pode ser representada como um problema de otimização quadrática. O algoritmo pode ser estendido para problemas não-lineares com o uso de funções *kernel*. Esse tipo de função aumenta, de forma não linear, a dimensionalidade dos dados. Uma função *kernel* pode mapear as instâncias de tal maneira que as classes se tornem linearmente separáveis após a sua aplicação. Problemas de classificação não binários (multi classes) podem ser decompostos em problemas binários com a indução de diferentes SVMs, usando por exemplo matrizes de código.

2.5. Árvore de Decisão

Segundo [Mitchell 1997], Árvores de Decisão é um dos métodos mais largamente utilizados para o processo indutivo. Esse método classifica as instâncias organizando-as em forma de árvore, do nó raiz até os nós folhas, cada nó especifica um teste de cada atributo da instância, cada ramo de um nó corresponde a um possível valor para o atributo, e a árvore é construída com base nas instâncias. As instâncias são testadas a partir do nó raiz, após criar um ramo para o valor do atributo, o processo se repete para este ramo criado, sendo considerado como um novo nó raiz na árvore. Uma medida muito utilizada para auxiliar a determinar e medir a pureza e impureza de um subconjunto é a entropia, que procura dizer o quão diferentes ou iguais elementos são entre si, com valores entre 0 e 1.

2.6. *Wikipedia e Simple Wikipedia*

A *Wikipedia*[1], segundo a própria definição disponível no site oficial, é um projeto de enciclopédia multilíngue de licença livre, escrita de maneira colaborativa e sem fins lucrativos. Possui um acervo de definições, palavras em grande escala, cada uma destas definições possui bastante informação relevante sobre o escopo do assunto, como contexto histórico, origem, entre outros aspectos. Segundo informações oficiais do site, em novembro de 2020 a Wikipédia em português possuía 1.047.992 artigos válidos, páginas de domínio principal. Por este motivo foi decidido por utilizar um *dataset* relacionado a esta enciclopédia devido a grande quantidade de dados e informações detalhadas de maneira completa que podem ser perfeitamente utilizadas no escopo do trabalho. A *Simple Wikipedia*[2] é uma versão da *Wikipedia* escrita apenas em inglês simplificado, em novembro de 2020 possuía 170 mil artigos, sendo a 50ª maior *Wikipedia*. Seu diferencial é utilizar sentenças curtas, palavras e gramáticas mais simples do que a *Wikipedia* completa. Seu *dataset* também será utilizado no objetivo do trabalho para treinar os algoritmos em representar a classe de textos com conteúdo mais fácil de compreender, por isso a escolha desse *dataset*.

3. Metodologia

A metodologia deste trabalho é exemplificada pelo fluxograma e sequência de atividades da Figura 2, porém para melhor entendimento, todas as etapas são descritas a seguir.

Para início da construção do protótipo foram necessárias muitas pesquisas, comparações e seguindo a indicação do orientador deste trabalho, chegou-se à conclusão da utilização dos datasets *Wikipedia* e *Simple Wikipédia*, a diferença entre eles permite com êxito aos algoritmos diferenciar uma definição complexa de uma simples. Após definidos os *datasets* é necessário realizar o tratamento, pré-processamento dos dados, com o objetivo de modificá-la para poder ser utilizada corretamente. Neste sentido, as bibliotecas citadas na Figura serão úteis para a realização do PLN.

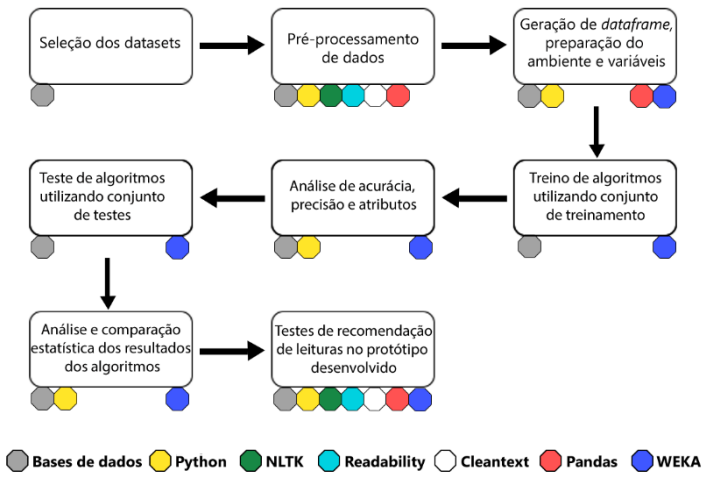

Figura 1. Diagrama de fluxo de trabalho. [Autoria própria]

[1] Site: https://www.wikipedia.org
[2] Site: https://simple.wikipedia.org

Após deixar o *dataset* preparado para ser processado, é necessário configurar o ambiente *Weka* e as variáveis dos algoritmos de classificação J48, SMO, e *Naive Bayes*, porém somente o parâmetro chamado número mínimo de instâncias por folha do algoritmo J48 é alterado, todos os demais mantém o padrão. Com este aspecto devidamente organizado, é possível realizar o treinamento dos algoritmos de classificação utilizando dos *datasets* no ambiente *Weka*. Conforme o processo de treinamento é finalizado, é possível obter estatísticas e a acurácia de cada um dos algoritmos, com estes dados é possível realizar a análise comparativa da eficiência dos algoritmos neste objetivo de classificação.

As etapas de treinamento dos algoritmos, análise dos *datasets* e resultados são aplicadas pelo ambiente Weka devido a ser uma ferramenta completa que já possui os algoritmos escolhidos para este trabalho implementados, sendo necessário somente alterar o parâmetro citado. Na etapa final com os algoritmos devidamente treinados e funcionais no objetivo de classificação, é possível utilizá-los como protótipo para os usuários, este protótipo terá como entrada outros *datasets* no mesmo formato apresentado anteriormente, e apresentará como saída, em forma de matriz de confusão, a classe da instancia, ou seja, o nível de legibilidade do texto presente no *dataset*, além da acurácia do algoritmo selecionado.

4. Resultados

Nesta seção serão apresentados as etapas e os resultados do desenvolvimento deste trabalho.

4.1. Pré-processamento

Os *datasets* utilizados são partes menores dos *datasets Wikipedia*[3] e *Simple Wikipedia*[4]. Possuem quantidades diferentes de artigos, sendo 19.555 presentes no dataset *Wikipedia*, e 49.754 no *dataset Simple Wikipedia*, porém estas quantidades foram redimensionadas para que ambos *datasets* tenham o mesmo número de instâncias. Após o redimensionamento as *datasets* ficaram com 19.544 instâncias para cada classe.

O *dataset Wikipedia* utilizado para este trabalho apresenta originalmente um formato e estrutura com diversas *tags* de Linguagem de Marcação de Hipertexto, o que dificulta que este seja aplicado ou utilizado diretamente no processo de extração de características textuais e obtenção das métricas, além dos artigos estarem contidos em um único arquivo, sendo necessário a separação de cada artigo em seu respectivo arquivo nomeado. Já o *dataset Simple Wikipedia* foi encontrado em formato de texto plano, ou seja, sem marcadores de linguagens de hipertexto, porém também todo reunida em um único arquivo de texto. Analisando estas condições foram desenvolvidos *scripts* que respectivamente modelam e modificam ambos em seus devidos formatos para a construção de um *dataset* de treinamento, os métodos *remove_stopwords* para retirar as palavras vazias das cadeias de caracteres, *tokenization* para separar as palavras em unidades e o *clean* do pacote *Cleantext* para modificar, retirar e substituir todos os conteúdos presentes nos textos que não são úteis para a classificação, por exemplo, remoção de e-mail, número de celular e substituindo-os por tokens representativas, além

[3]Site:http://wikipedia.c3sl.ufpr.br/enwiki/20210620/enwiki-20210620-pages-meta-current1.xml-p1p41242.bz2
[4]Repositório: https://github.com/LGDoor/Dump-of-Simple-English-Wiki

da mudança nas palavras de caixa alta para caixa baixa. A Figura 3 apresenta um exemplo de artigo dos *datasets* processado.

```
("['a', 'zoological', 'garden', ',', 'zoological', 'park', ',', 'zoo', "
 "'place', 'different', 'species', '(', 'types', ')', 'animals', 'kept', "
 "'held', 'bad', 'conditions', '.', 'they', 'kept', 'small', 'cages', ',', "
 "'bored', 'sick', '.']")
```

Figura 2. Artigo de exemplo do dataset depois do processamento do texto para texto plano. [Autoria própria]

Para utilizar algoritmos para extração de características e métricas textuais também foi desenvolvido um *script* que utiliza dos textos processados e os enviam como parâmetro para o método *readability*, este obtém as diversas métricas e características textuais e retorna um dicionário. O *dataset* para treinamento dos algoritmos é construído com atributos normalizados em torno das métricas obtidas, nome dos artigos, resultados de fórmulas de legibilidade e nome do *dataset* de origem do artigo. A estrutura do *dataset* é apresentado conforme a Figura 4.

```
characters_per_word,syll_per_word,words_per_sentence,...,classes
0.17639876715210306,0.2519110021059938,0.030066200809121,...,enwikipedia
0.15367011865406388,0.220557500119716,0.04067058967757754,...,enwikipedia
0.15064364896907834,0.2194588771288639,0.13816354051734708,...,enwikipedia
```

Figura 3. Exemplo de formato do dataset de treinamento. [Autoria própria]

4.2. Análise do *dataset*

Com o *dataset* gerado pode ser feita a análise sobre as instâncias e valores. Através de gráficos representativos do *dataset* é possível observar algumas características das instâncias de diferentes classes. A Figura 5 apresenta graficamente alguns dos atributos do *dataset* com valores normalizados entre 0 e 1, listados como entre os mais importantes, segundo algoritmos de seleção de atributos do ambiente WEKA para a classificação das instâncias.

No *dataset Simple Wikipedia* quase toda a frequência normalizada de palavras complexas se concentra próxima ao valor zero, assim não presentes tantas palavras complexas que é seu objetivo, enquanto o *dataset Wikipedia* já tem as frequências mais distribuídas, pois o objetivo da base é conter definições mais complexas dos artigos presentes. A frequência normalizada de palavras longas, apresenta-se de maneira bem similar ao comportamento da frequência de palavras complexas. Palavras muito longas podem atrapalhar a leitura do usuário sobre um texto qualquer, ciente deste aspecto o *dataset Simple Wikipedia* foi construído visando não utilizar palavras longas e por este motivo a frequência fica bem próxima do valor zero, enquanto para o *dataset Wikipedia* esta frequência normalizada é mais distribuída. O índice Miyazaki [Greenfield 2004] é semelhante aos índices *Kincaid Grade Level* [Kincaid et al. 1975] e *Flesch Reading Ease* [Flesch 1948]. O valor normalizado desse índice, para os artigos do *dataset Simple Wikipedia* apresentou maior frequência de valores próximos a um, ou seja, leituras mais fáceis do que o *dataset Wikipedia*.

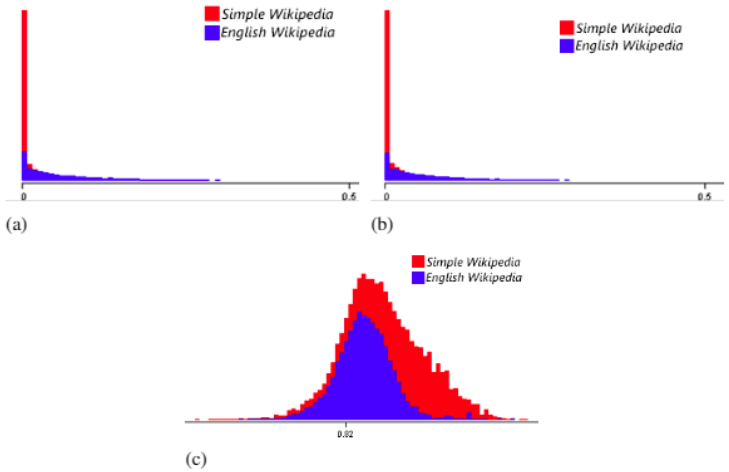

Figura 4. Gráfico (a) frequência de palavras complexas. Gráfico (b) frequência de palavras longas. Gráfico (c) índice de Miyazaki. [Autoria própria]

4.3. Aprendizado de máquina

Os algoritmos de aprendizado de máquina são aplicados no *dataset* através do ambiente WEKA na versão 3.8.5. Para este trabalho são aplicados os algoritmos Máquina de Vetor Suporte (*LibSVM*), *Naive Bayes* e árvore de decisão (J48). Todos os algoritmos utilizaram parâmetros padrões do programa, exceto pela árvore de decisão J48, a qual o parâmetro chamado número mínimo de instâncias por folha foi configurado com valor igual a 800, desta forma, realiza-se uma pré-poda com o objetivo de gerar uma árvore menor. Através do programa é possível acessar a estrutura visual da árvore formada pelo algoritmo de Árvore de decisão, com base nisso a interpretação sobre as instâncias classificadas se torna um processo mais visível, a estrutura resultante da construção da árvore para as instâncias do algoritmo é demonstrada pela Figura 6.

A Tabela 1 apresenta os resultados do treinamento dos algoritmos de classificação.

Tabela 1 – Acurácias obtidas pelos algoritmos de treinamento. [Autoria própria]

Algoritmo	Categoria	Acurácia	Método de reamostragem
J48	Árvore de decisão	94.17%	Divisão de porcentagem
J48	Árvore de decisão	90.12%	Validação cruzada
LibSVM	Máquina de vetor suporte	87.67%	Divisão de porcentagem
LibSVM	Máquina de vetor suporte	87.33%	Validação cruzada
NaiveBayes	*Naive Bayes*	85.63%	Divisão de porcentagem
NaiveBayes	*Naive Bayes*	85.34%	Validação cruzada

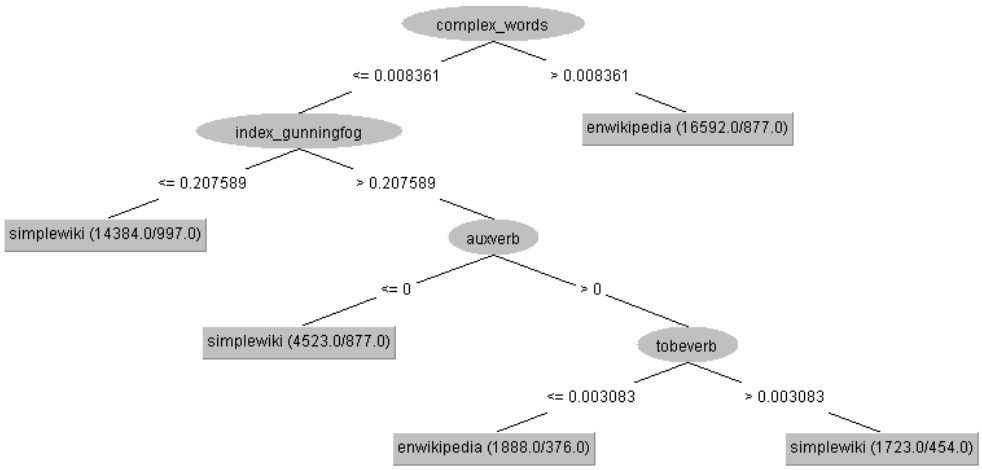

Figura 5. Visualização da estrutura da árvore construída para o *dataset*.

A figura demonstra que, para este algoritmo, grandes quantidades de instâncias podem primeiramente ser classificadas como pertencentes a classe *Wikipedia*, ou seja, um texto com maior índice de dificuldade de leitura, levando em consideração a frequência de palavras complexas presentes no texto. Já o índice de *Gunning Fog* pode ser um separador de grandes quantidades de instâncias pertencentes a classe *Simple Wikipedia*, textos com menor dificuldade de leitura. A ausência de verbos auxiliares e a maior frequência de *to be verbs* demonstraram serem aspectos de artigos do *dataset Simple Wikipedia*. *To be verbs* são uma categoria de verbos mais simples da língua inglesa, e, portanto, a maior frequência deste tipo de verbo no *dataset Simple Wikipedia* era esperada devido a sua característica.

Este algoritmo também, de todos os testados, foi o que apresentou maior acurácia em ambos os métodos de reamostragem. Em segundo lugar na avaliação de acurácia dos algoritmos, o SVM, apresentou também um desempenho condizente com trabalhos da área, como o de [Scarton and Aluísio 2010], no qual o algoritmo havia demonstrado a melhor acurácia e resultado comparado aos demais algoritmos. A escolha de diferentes características, principalmente no objetivo de classificação textuais em níveis de dificuldade de leitura, pode ser um fator muito impactante nos resultados obtidos pelos algoritmos. A escolha de um algoritmo probabilístico simples baseado no teorema de *Bayes* foi proposital para a análise da dificuldade do objetivo de classificação. Mesmo um algoritmo que parte ingenuamente de suposições de fortes independências entre as características, consegue alcançar uma acurácia relevante no objetivo de classificação com as características selecionadas e utilizadas neste trabalho.

5. Conclusão

Nesse trabalho foi apresentado a classificação de textos em níveis de dificuldade de leitura através da utilização de algoritmos clássicos de aprendizado de máquina e bases de dados gratuitas e disponíveis na Internet que continuam a serem atualizadas constantemente como meio de treinamento. O algoritmo de Árvore de decisão apresentou o melhor resultado, utilizando parâmetros padrões do programa WEKA alcançou acurácia de 94.17%, apresentando os atributos frequência de palavras complexas, frequência de verbos auxiliares, *to be verb* e índice de Gunning Fog como importantes para a diferenciação das classes.

Este classificador pode ser utilizado como base para diferentes finalidades, por exemplo, sistemas de recomendação de conteúdo para aprendizes da língua inglesa como segunda língua, uso individual para pessoas com baixo letramento no objetivo de desenvolver a leitura em diferentes contextos, além de protótipo de ferramenta de leitura para auxílio de pessoas com distúrbios cognitivos que afetam a fala, nos aspectos de expressão, entendimento da linguagem e aprendizagem da leitura, como a afasia e dislexia. Possibilita e abre espaço para criação de diversas outras ferramentas e aplicações que auxiliam usuários, complementando ferramentas gratuitas e disponíveis na área de legibilidade e inteligibilidade de textos através do Processamento de Língua Natural.

6. Agradecimentos

Agradecimentos aos meus professores, amigos e companheiros de curso, meu orientador do Trabalho de Conclusão do Curso Arnaldo Cândido Júnior, por terem me auxiliado na realização do trabalho, ter fornecido informações indispensáveis para o desenvolvimento. Agradecimento à Universidade Tecnológica Federal do Paraná campus Medianeira por todo suporte deste trabalho.

Referências

BAKER, W. English as a lingua franca and intercultural communication. The Routledge Handbook of English as a Lingua Franca, v. 17, n. 3, p. 25–36, 2018.

British Council. Demandas de aprendizagem de inglês no Brasil. São Paulo, 2014.

COVINGTON, M. A. Natural Language Processing for Prolog Programmers. [S.l.]:Prentice-hall, Inc., 2013.

FLESCH, R. A New Readability Yardstick. [s.n.], 1948. Disponível em: <https://books.google.com.br/books?id=C0xkNQEACAAJ>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. [S.l.], 2016.

GREENFIELD, J. Readability Formulas For EFL.JALT Journal, v. 26, n. 1, p. 5, 2004.

KINCAID, J. P. J. et al. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel (No.RBR-8-75). Naval Technical Training Command Millington TN Research Branch. Naval Technical Training Command Millington TN Research Branch, 1975. Disponível em: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA006655>.

MELITZ, J. English as a global language. In: The Palgrave Handbook of Economics and Language. [S.l.]: Palgrave Macmillan, 2016. p. 583–615. ISBN 9781137325051.

MITCHELL, T. M. T. M. Machine Learning. [S.l.: s.n.], 1997. 414 p. ISBN 0070428077.

RUSSELL, S.; NORVIG, P. A AI I PRENTICE HALL SERIES IN ARTIFICIAL INTELLIGENCE. [S.l.], 2013.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: Adaptando as métricas do Coh-Metrix para o Português. Linguamatica, v. 2, p. 45–62, 2010. ISSN 16470818.