UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

RAFAEL FELIPE SANDRONI DIAS

Caracterização autoral a partir de textos utilizando redes neurais artificiais

São Paulo
2019

RAFAEL FELIPE SANDRONI DIAS

Caracterização autoral a partir de textos utilizando redes neurais artificiais

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 11 de outubro de 2019. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Ivandré Paraboni

São Paulo

2019

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Dias, Rafael Felipe Sandroni
Caracterização autoral a partir de textos utilizando redes neurais artificiais / Rafael Felipe Sandroni Dias ; orientador, Ivandré Paraboni. – 2019.
103 f. : il

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Reconhecimento de padrões. 2. Inteligência artificial. 3. Redes neurais. 4. Reconhecimento de texto. I. Paraboni, Ivandré, orient. II. Título

CDD 22.ed.– 006.4

Dissertação de autoria de Rafael Felipe Sandroni Dias, sob o título **“ Caracterização autoral a partir de textos utilizando redes neurais artificiais ”**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 11 de outubro de 2019 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Universidade de São Paulo

Presidente

Prof. Dra. Sarajane Marques Peres

Universidade de São Paulo

Prof. Dra. Helena de Medeiros Caseli

Universidade Federal de São Carlos

Prof. Dr. Marcelo de Souza Lauretto

Universidade de São Paulo

Agradecimentos

Agradeço à minha família, que sempre me apoiou e me deu o suporte necessário para o desenvolvimento deste trabalho. A todos os professores que me ajudaram promovendo ideias e conhecimento. Aos colegas de laboratório, que de alguma forma, colaboraram com discussões valiosas e feedbacks. Ao meu orientador, pelo conhecimento compartilhado e papel crucial para a finalização deste trabalho.

Resumo

DIAS, Rafael Felipe Sandroni. **Caracterização autoral a partir de textos utilizando redes neurais artificiais**. 2019. 103 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2019.

A caracterização autoral (CA) é uma tarefa computacional de reconhecimento de características de autores de textos com base em seus padrões linguísticos. O uso de modelos computacionais de CA permite inferir características sociais a partir do texto, mesmo que os autores não escolham conscientemente colocar indicadores dessas características no texto. A tarefa de CA pode ser importante para diversas aplicações práticas, tais como análise forense e marketing. Abordagens tradicionais de CA muitas vezes utilizam conhecimento linguístico, que exige conhecimento prévio e demanda esforço manual para extração de características. Recentemente, o uso de redes neurais artificiais têm demonstrado resultado satisfatório em problemas de processamento de linguagem natural (PLN), entretanto, para caracterização autoral, apresenta um nível variado de sucesso. Este trabalho tem o objetivo de organizar, definir e explorar diversas tarefas de caracterização autoral a partir de córpus textuais, abrangendo três idiomas (i.e., português, inglês e espanhol) e quatro domínios textuais (i.e., redes sociais, questionários, SMS e blogs). Foram propostos seis modelos baseados em redes neurais e *Word Embeddings*, comparando-se com sistemas de *baseline* utilizando regressão logística e TF-IDF. Os resultados dos modelos de *Long Short Term Memory* (LSTM) *with self-attention* e *Convolutional Neural Network* (CNN) sugerem que tais técnicas apresentam desempenho superior ao *baseline* quando córpus grandes são utilizados. Os modelos de LSTM *with self-attention* baseados em representação de *Word Embeddings* e *Char* apresentam desempenho superior ao estado da arte da competição PAN-CLEF 2013.

Palavras-chaves: Caracterização autoral. Redes Neurais Artificiais. Word Embeddings.

Abstract

DIAS, Rafael Felipe Sandroni. **Author Profiling from texts using artificial neural networks**. 2019. 103 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2019.

Author Profiling (AP) is a computational task of recognizing the characteristics of text authors based on their linguistic patterns. The use of computer computational models allows us to infer social characteristics from the text, even if the authors do not consciously choose to place indicators of these characteristics in the text. The AP task can be important for many practical applications, such as forensic analysis and marketing. Traditional AP approaches often use language knowledge, which requires prior knowledge and requires manual effort to extract features. Recently, the use of artificial neural networks has shown satisfactory results in natural language processing (NLP) problems, however, for author profiling, presents a varied level of success. This paper aims to organize, define and explore various authorial characterization tasks from the textual corpus considered, covering three languages (i.e, Portuguese, English and Spanish) and four textual domains (ie, social networks, questionnaires, SMS and blogs) . Six models based on neural networks and Word Embeddings were proposed, compared with baseline systems using logistic regression and TF-IDF. The results suggest that the *Long Short Term Memory with self-attention* and *Convolutional Neural Network* models outperform baseline system in larger volume corpus. The LSTM *with self-attention* model based on *Word Embeddings* and *Char* text representation outperform the state-of-the-art PAN-CLEF 2013 competition.

Keywords: Author Profiling. Artificial Neural Networks. Word Embeddings.

Lista de figuras

Lista de quadros

Lista de tabelas

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A caracterização autoral (CA) (do inglês, *Author Profiling*) é uma tarefa computacional de reconhecimento de características sociais de autores de textos com base em seus padrões linguísticos. Estes padrões são compartilhados por autores que apresentam características sociais semelhantes, como a faixa etária, gênero ou grau de escolaridade (NGUYEN et al., 2016).

Com o surgimento de mídias sociais, cada vez mais pessoas produzem alguma postagem online, composta principalmente por conteúdo de imagem e texto. Isso significa que essas pessoas podem ser consideradas autores e, portanto, é possível identificá-las como tal. O uso de modelos computacionais de CA permite inferir características sociais a partir do texto, mesmo que os autores não escolham conscientemente colocar indicadores dessas características no texto. A pesquisa de CA tem se tornado um campo amplo e de crescente interesse (RANGEL; ROSSO, 2019).

A tarefa de CA pode ser importante para diversas aplicações práticas. Pela perspectiva de análise forense, identificar padrões linguísticos de autores de mensagens suspeitas e reconhecer características sociais desses autores com base em análises de textos pode ajudar a considerar novos suspeitos em investigações criminais (RANGEL et al., 2017). Além disso, a tarefa de CA pode também auxiliar na identificação de perfis de robôs associados à disseminação de notícias falsas em redes sociais (RANGEL; ROSSO, 2019), e de perfis falsos usados no recrutamento terrorista e *Cyberbulling*, como sugerido em Russell e Miller (1977). Pelo ponto de vista de marketing, empresas podem também ter interesse em segmentar seu público alvo por características sociais. Tais aplicações, geralmente, são empregadas em diversos domínios textuais e idiomas, e tarefas de CA mais específicas podem refinar o desempenho dessas aplicações.

Esforços crescentes da competição PAN-CLEF (RANGEL et al., 2017) têm reunido pesquisadores de diversas áreas e contribuído significativamente para o desenvolvimento de novos modelos de CA. Entretanto, a tarefa de CA configura um problema ainda não totalmente resolvido (NGUYEN et al., 2016). Os estudos de CA, de modo geral, tendem a utilizar abordagens de extração de características dependentes do domínio e idioma do texto, tornando tais modelos difíceis de generalizar para outros tipos textuais (e.g., redes sociais, blogs, mensagens) e idiomas (GOPINATHAN; BERG, 2017). Tais abordagens exigem

um elevado nível de trabalho manual para computação de características e conhecimento linguístico prévio. Além disso, observa-se que estes estudos são frequentemente voltados para o idioma inglês e limitados às tarefas de caracterização de faixa etária e gênero (RANGEL et al., 2017; RANGEL et al., 2018). Outras tarefas, como caracterização de grau de educação e grau de religiosidade, permanecem pouco exploradas.

Estudos recentes em diversas áreas de processamento de língua natural (PLN) têm obtido resultados satisfatórios com uso de redes neurais recorrentes (RNNs) (CHUNG et al., 2014; LAI et al., 2015) e redes neurais de convolução (CNNs) (ZHANG; ZHAO; LECUN, 2015a; CONNEAU et al., 2016; KIM, 2014a). No caso da CA, estudos iniciais deste tipo demonstram um nível variado de sucesso (GOPINATHAN; BERG, 2017; SIERRA et al., 2017; TAKAHASHI et al., 2018).

1.1 Objetivo

O objetivo geral deste trabalho é desenvolver modelos de aprendizado de máquina (AM) supervisionado baseados em RNAs e representação distribuída de palavras (i.e., *word embeddings*) que apresentam resultados superiores do que modelos de CA utilizando regressão logística e TF-IDF (do inglês, *Term Frequency - Inverse Document Frequency*), para o reconhecimento de múltiplas tarefas de CA utilizando córpus de diversos domínios e idiomas.

De forma mais específica, considera-se desenvolver modelos de CA para a caracterização de gênero, faixa etária, grau de escolaridade, grau de religiosidade, formação em TI e posição política, com base em textos rotulados provenientes de córpus textuais nos idiomas português, inglês e espanhol. Os córpus a serem considerados são: PAN-CLEF 2013 (RANGEL et al., 2013), *The Blog Authorship* (SCHLER et al., 2006), b5-post (RAMOS et al., 2018), BRMoral (SANTOS; PARABONI, 2019) e BlogSet-BR (SANTOS; WOLOSZYN; VIEIRA, 2018).

1.2 Hipótese

Este estudo considera a hipótese de que o uso de técnicas de RNAs combinadas com representação distribuída de palavras proporcionará resultados superiores aos de

modelos tradicionais de CA que não utilizam essas técnicas. Esta hipótese será testada comparando-se os modelos propostos com sistemas de *baseline* que se façam pertinentes usando as métricas de avaliação de medida F (i.e., F_1 *score*) e acurácia. Espera-se que, com um resultado positivo deste estudo, seja demonstrado que RNAs e representação distribuída de palavras são uma alternativa adequada para implementação de modelos de CA independentes do domínio e idioma do texto.

As contribuições previstas para este trabalho são: (1) a organização de diversos córpus e tarefas de CA, (2) modelos baseados em RNAs para estas tarefas e (3) resultados de referência para futuros estudos desta área.

1.3 Organização do documento

O restante deste documento apresenta uma pesquisa em nível de mestrado tratando do problema de CA baseada em diversos córpus e idiomas. Os capítulos seguintes apresentam as revisões bibliográficas de conceitos fundamentais (Capítulo 2), discussões acerca de estudos correlatos encontrados na literatura (Capítulo 3), estudos exploratórios e resultados iniciais (Capítulo 4), organização dos dados (Capítulo 5), modelos propostos (Capítulo 6), avaliação dos resultados (Capítulo 7) e conclusão (Capítulo 8).

2 Conceitos fundamentais

Este capítulo apresenta os conceitos fundamentais, relacionados ao escopo desta pesquisa, acerca das técnicas usadas na área de caracterização autoral.

2.1 Caracterização autoral

A caracterização autoral (CA) é uma tarefa computacional de reconhecimento de características de autores de textos com base em padrões linguísticos. Estes padrões são compartilhados por autores que apresentam características sociais semelhantes, como a faixa etária, gênero ou grau de escolaridade (NGUYEN et al., 2016).

O conceito de dialetos sociais é similar ao conceito de dialetos regionais. Os dialetos regionais são variações linguísticas baseadas na geografia do autor, e dialetos sociais são variações linguísticas baseadas em traços compartilhados por grupos sociais. Por meio dos dialetos sociais, é possível identificar grupos de autores de textos de acordo com o gênero, faixa etária, grau de escolaridade etc.

Tendo em vista que autores podem ajustar sua comunicação e padrão linguístico, de forma consciente ou inconsciente, para um ambiente específico, há discussões acerca do uso de variáveis como gênero e idade serem consideradas características sociais do autor, que podem ser mutáveis, ao invés de características biológicas, que são informações imutáveis. Desta forma, também há discussões acerca de variações linguísticas (i.e., apresentar um posicionamento formal ou informal, por exemplo) que refletem o assunto ou público que o autor envolvido (REDDY; VARDHAN; REDDY, 2017).

Do ponto de vista computacional, a análise de variações linguísticas de autores a partir de textos pode ser subdividida em problemas onde há textos de um mesmo autor em diversos tipos textuais (e.g., redes sociais, e-mails, blogs), até problemas de processamento multilíngue nos quais um mesmo autor pode empregar diversos idiomas em um mesmo texto. Entretanto, a automação da predição de características de autores é considerada uma tarefa não-trivial. Estudos da área têm comparado o desempenho de seres humanos contra sistemas automatizados e constataram, por exemplo, que estes sistemas desempenham melhor na identificação de gênero e idade de pessoas com base em suas publicações no Twitter (BURGER et al., 2011; NGUYEN et al., 2014).

A predição de gênero e idade são as tarefas de CA mais comuns encontradas na literatura, com diversos estudos que tratam estes problemas a partir de conjuntos de dados que trazem essas informações rotuladas e disponibilizadas publicamente (Guimarães et al., 2017; KIM et al., 2017). Por outro lado, outras tarefas como a predição de variação de idioma (SIERRA et al., 2017), renda (FLEKOVA; PREOTIUC-PIETRO; UNGAR, 2016) ou personalidade (GONZÁLEZ-GALLARDO et al., 2015) são menos comuns, possivelmente porque são informações mais difíceis de obter, e muitas vezes exigindo extensas coletas de dados.

Entre as tarefas de CA mais comuns, a tarefa de predição de idade é muitas vezes tratada como um problema de classificação multiclasse, e em alguns estudos como problema de classificação binária (Guimarães et al., 2017) e regressão (SAP et al., 2014; FLEKOVA; PREOTIUC-PIETRO; UNGAR, 2016). A tarefa de gênero é geralmente tratada como um problema de classificação binária.

2.1.1 Competições PAN-CLEF

A CA tem sido tema de uma série de competições de análises forenses de textos digitais, PAN-CLEF (RANGEL et al., 2017). A edição de 2013 (RANGEL et al., 2013) apresentou as tarefas de predição de idade e gênero, junto com conjuntos de dados de redes sociais nos idiomas inglês e espanhol. Na edição de 2014 (RANGEL et al., 2014), mantiveram-se as tarefas de predição de idade e gênero, e em adicional, um conjunto de dados em quatro tipos textuais: publicações de blogs, Twitter, Facebook e avaliações de hotéis. Exceto para o conjunto de avaliações de hotéis, composto por textos apenas em inglês, todos conjuntos foram apresentados nos idiomas inglês e espanhol. A edição de 2015 (RANGEL et al., 2015), similar às edições anteriores, adicionou a tarefa de predição de personalidade e conjuntos de dados com dois novos idiomas (italiano e holandês) para as tarefas de predição de idade e personalidade. A edição de 2016 (RANGEL et al., 2016) considerou a investigação dos efeitos da avaliação de modelos treinados em um tipo textual (e.g., Twitter, blogs e redes sociais) e avaliados em outro para predição de idade e gênero, nos idiomas inglês, espanhol e holandês. Na edição de 2017 (RANGEL et al., 2017) foi introduzida a tarefa de predição de variação de idioma, junto com predição de gênero, nos idiomas inglês, espanhol, português e árabe. Por fim, a edição 2018[1] abordou a tarefa de

[1] https://pan.webis.de/clef18/pan18-web/author-profiling.html

predição de gênero como um problema de classificação de imagens e textos a partir de fotos e *tweets* do Twitter, nos idiomas inglês, espanhol e árabe.

A competição PAN-CLEF tem um papel importante na área de CA pois traz avanços significativos aos problemas explorados, e crescente número de pesquisadores e profissionais participantes de discussões da área. Em todas as suas edições, a PAN-CLEF trouxe abordagens relevantes para problemas que impactam a vida de pessoas em ambientes digitais. Além disso, os conjuntos de dados disponibilizados permitem a replicação de trabalhos por outros pesquisadores, colaborando com o avanço de novas fronteiras do conhecimento na área de CA.

2.1.2 Tipos de conhecimentos utilizados na CA

Os estudos de CA adotam diferentes tipos de conhecimento e estratégias para extrair características e padrões linguísticos em textos. Estes conhecimentos são, muitas vezes, combinações de estratégias de extração de características dependentes de domínio e idioma do texto, e técnicas de representação textual. Em Reddy, Reddy e Vardhan (2016) são apresentados os principais tipos de características usados em CA. De modo geral, são características baseadas em sete tipos de informações: de palavras, caracteres, sintáticas, documentos, tópicos, nível de legibilidade e recuperação de informação (IR) (do inglês, *Information Retrieval*), e utilizam-se de técnicas estatísticas, regras morfológicas e sintáticas, e vocabulários.

- **Palavras**: Utilizam informações de palavras no texto, é a estratégia mais popular em CA, e adotam principalmente técnicas de contagem, frequência de palavras e conjuntos de palavras, i.e., BoW (do inglês, *Bag of Words*), TF-IDF e n-gramas de palavras. Em alguns casos, são consideradas informações de vocabulários externos (i.e., *stopwords*, lematização, gírias, etc.) e técnicas de análise de sentimento para rotular, transformar ou remover palavras (MEINA et al., 2013a).
- **Caracteres**: Utilizam informações de nível de caracteres no texto, e adotam técnicas contagem e frequência de caracteres (i.e., BoW e TF-IDF) e, principalmente, de conjuntos de caracteres (i.e., n-gramas de caracteres) (BASILE et al., 2017).
- **Características sintáticas**: são informações de estrutura gramatical e utilizam técnicas para extração de regras gramaticais e técnicas estatísticas baseadas em es-

trutura morfológica e gramatical (i.e., frequência de pronomes, substantivos, palavras no singular e plural, pontuação, erros gramaticais, etc.) (GONZÁLEZ-GALLARDO et al., 2015).

- **Características baseadas em documentos**: são informações estruturais de documentos de domínios específicos e utilizam técnicas estatísticas de contagem de parágrafos e sentenças, além de técnicas especializadas em domínios, tais como etiquetas para identificar menções, *hashtags*, *retweets*, URLS, *emoticons*, dicionários de gírias e elementos HTML (MARTINC et al., 2017; SAP et al., 2014).
- **Características baseadas em tópico**: são informações de tópicos específicos (e.g., esportes, política, relacionamento, etc.) e utilizam técnicas para extração de palavras chaves, vocabulários psicolinguísticos e técnicas de análise de sentimento (ISBISTER; KAATI; COHEN, 2017)
- **Características de nível de legibilidade**: são informações de complexidade de compreensão do texto e utilizam técnicas especializadas em extração e anotação de informações de legibilidade, tais como *Flesch Knkaid Grade Level* e *Flesch Reading Ease* (KINCAID et al., 1975; WEREN et al., 2014).
- **Características de recuperação de informação**: são informações de medidas de similaridades entre textos e, geralmente, utiliza-se as medidas cosseno e Okapi BM25 (MANNING; RAGHAVAN; SCHüTZE, 2008). Em Weren et al. (2014) são discutidas 30 abordagens de características baseadas em similaridades para CA.
- **Características de segunda ordem**: É um tipo de conhecimento específico para CA, baseado na relação entre termos, documentos e características de autores (i.e., classes de tarefas de CA). Utilizam vetores de representação de termos para estimar a relação entre o uso de termos e características de autores, e representação de documentos, computando os termos de cada documento, para estimar a relação geral de documentos e características de autores.

Além das características discutidas por Reddy, Reddy e Vardhan (2016), há modelos baseados em vocabulários pré-treinados e dependentes de idioma: etiquetas *Part-of-Speech* (POS) e dicionário *Linguistic Inquiry and Word Count* (LIWC) (ISBISTER; KAATI; COHEN, 2017; VOLLENBROEK et al., 2016; REDDY; VARDHAN; REDDY, 2017).

Os modelos de POS são baseados em ferramentas que analisam textos e fazem anotações de etiquetas morfológicas e sintáticas para cada termo. Essas etiquetas categori-

zam palavras com base na estrutura gramatical e posição de termos adjacentes em um documento. Este modelo depende de um conjunto de etiquetas pré-treinadas para cada idioma. Entretanto, há discussões acerca de regras multilíngue em Nivre e Fang (2017).

O modelo de conhecimento psicolinguístico do dicionário LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001) também é baseado em ferramentas de análise de textos e faz a categorização de termos em significados psicolinguísticos, em relação à estrutura gramatical (i.e., artigos, pronomes etc.) e em relação ao conteúdo ou tópico específico (i.e., termos relacionados a sucesso, morte, etc.). O dicionário LIWC é atualmente traduzido em mais de 20 idiomas, e possui uma versão em português discutida em Filho, Aluísio e Pardo (2013).

2.2 Métodos de representação textual

Além das características dependentes de domínio ou idioma discutidas na seção anterior, modelos computacionais CA utilizam também representação vetorial de textos. Além disso, é realizada a extração de informações semânticas e sintáticas sem dependência de conhecimento linguístico, domínio e idioma. Para o propósito deste capítulo, iremos discutir dois modelos: tradicionais e representação distribuída de palavras.

2.2.1 Modelos tradicionais

O modelo *Bag of Words* (BOW) adota a representação simbólica em vetores de números naturais $\vec{w}_d \in \mathbf{N}$ para contagem e frequência de palavras. Neste modelo, o documento é representado pelo vetor $\vec{w}$ de dimensão d igual ao total de palavras únicas no vocabulário do texto, contendo o número de ocorrências (ou frequência) de cada palavra e desconsiderando a estrutura gramatical e a ordem das palavras.

Apesar de utilizar uma representação textual simplificada, este modelo é amplamente utilizado na área de PLN justamente por sua simplicidade. Entretanto, uma das suas desvantagens é a sua representação esparsa, gerando modelos com alta dimensionalidade, por exemplo, considera-se um vocabulário de 40.000 palavras para representar um documento de 20 palavras, este modelo terá apenas 20 dimensões com valores não zeros (GOLDBERG; HIRST, 2017).

Modelos baseados em frequência TF-IDF (do inglês, *Term Frequency - Inverse Document Frequency*) adotam uma representação em vetores de números reais $\vec{w}_d \in \mathbf{R}$ e evidencia os termos mais raros no documento aplicando a medida TF-IDF para indicar a importância de um termo em relação ao documento. Um termo pode ser um *token*, palavra ou n-grama. Quanto maior o número de ocorrências de um termo no documento, maior é a medida TF-IDF deste termo.

O modelo de n-gramas, também similar ao modelo BoW, utiliza representação em vetores de sequências de n termos consecutivos. Estes termos podem ser palavras (n-gramas de palavras), caracteres (n-gramas de caracteres) ou POS (n-gramas de POS), onde n é o tamanho da sequências de termos. Quando temos $n = 1$ os termos podem ser descritos como unigramas, quando $n = 2$ são bigramas etc.

2.2.2 Representação distribuída de palavras (*Word embeddings*)

Tradicionalmente, modelos de representação textual baseados em contagem usam representação simbólica, onde cada palavra é representada por uma única dimensão no vetor e a dimensionalidade deste vetor é a mesma do número de palavras distintas do vocabulário. Além disso, as palavras são independentes umas das outras, mesmo que o significado destas palavras sejam similares.

Em contraste aos modelos tradicionais de representação textual, as redes neurais artificiais (RNA) são frequentemente utilizadas para extrair características de representação distribuída de palavras similares com base em contextos similares, e mapeá-las em vetores (GOLDBERG; HIRST, 2017). Cada palavra é incorporada em um espaço dimensional d e representada como um vetor $\vec{w}$ de números reais $\vec{w} \in \mathbf{R}_d$. O contexto da palavra é capturado nas diferentes dimensões d do vetor. Estas dimensões não são interpretáveis e dimensões específicas não necessariamente correspondem a significados específicos. Por essa natureza, um determinado significado pode ser capturado por diferentes combinações de dimensões, e uma determinada dimensão pode contribuir para capturar diferentes aspectos do significado de uma palavra (GOLDBERG; HIRST, 2017).

A dimensão é normalmente menor que o número de palavras no vocabulário. Por exemplo, em um vocabulário de 40.000 palavras, cada palavra pode ser representada por um vetor de 100 ou 200 dimensões. Baseado na hipótese distribucional de Harris (1954),

o significado de uma palavra pode ser dado pelo seu emprego em determinado contexto, uma palavra pode ser derivada a partir de sua distribuição do córpus e da agregação das palavras do contexto em que ela ocorre (GOLDBERG; HIRST, 2017).

Em Bengio et al. (2003) são apresentados modelos neurais de língua (do inglês, *Neural Natural Language Modeling* - NNLM) para capturar informações de representação a partir de informações de contexto. Estes modelos têm o objetivo de representar uma determinada palavra com base nas palavras que ocorrem em seu contexto. Mais recentemente, algumas ferramentas implementaram modelos neurais de língua baseados em Bengio et al. (2003). O algoritmo Word2vec (MIKOLOV et al., 2013a) apresenta a implementação de duas variações de arquiteturas para representação de contextos: CBOW e *Skipgram*, e funções de otimização *Negative Sampling* e *Hierarquical Softmax*.

A variação *Continuous Bag-of-Words* (CBOW) considera um contexto $\vec{C}$ de k palavras ($\vec{C_k}$), e define o vetor de contexto $\vec{C}_{1:k}$ como a soma dos vetores distribuídos de palavras deste contexto:

$$\vec{C} = \sum_{i=1}^{k} c_i \tag{1}$$

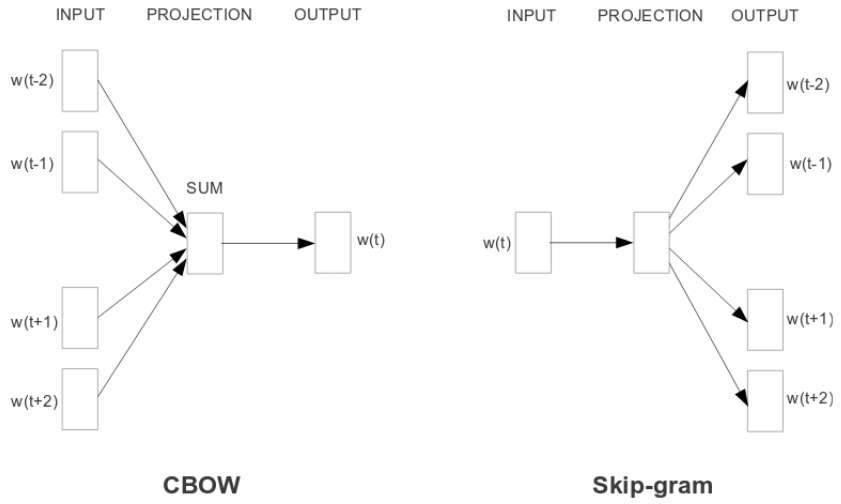
O CBOW possui semelhança com o modelo BOW pois não retém informações da ordem das palavras no contexto, apenas a presença de tais palavras dentro da uma janela de contexto (GOLDBERG; HIRST, 2017).

A variação *Skipgram* considera as palavras dentro de um mesmo contexto como independentes. Em uma janela de contexto de $\vec{C}$ de k palavras ($\vec{C}_{1:k}$), o *Skipgram* assume que as palavras no contexto $\vec{C}$ são independentes umas das outras, tratando-as como k diferentes contextos. O par de palavras w e contexto $\vec{C}_{i:k}$ será representado em D como k diferentes contextos, $(w : \vec{C_1})$,$(w : \vec{C_2})$...,$(w : \vec{C_k})$ (GOLDBERG; HIRST, 2017).

A figura 1 apresenta uma ilustração das arquiteturas CBOW e *Skipgram*: a primeira faz a predição da palavra atual com base no seu contexto, e a segunda faz a predição do contexto com base na palavra atual.

Além do algoritmo Word2Vec, outras ferramentas de representação distribuída de palavras usando modelos de língua neurais foram desenvolvidas. GloVe (PENNINGTON; SOCHER; MANNING, 2014) constrói uma matriz de contextos C de k palavras com base em informações de ocorrências de uma palavra $(k_1 : c_1)$ no contexto de outra palavra $(k_2 : c_2)$, definindo coocorrências entre palavras e contextos. Além deste, o FastText

Figura 1 – Arquiteturas do Word2Vec.



Fonte: (MIKOLOV et al., 2013a)

proposto em Joulin et al. (2016) realiza o treinamento da rede usando n-gramas de caracteres. O algoritmo é capaz de capturar informações mais detalhadas do contexto e apresenta tolerância a erros ortográficos. As variações de CBOW e *Skipgram* também são implementadas pelo FastText.

2.3 Métodos de aprendizado de máquina para CA

Tradicionalmente, modelos de CA fazem uso de métodos de aprendizado de máquina (AM) para classificação de textos. Entre estes métodos, máquinas de vetores de suporte (SVM) (do inglês, *Support Vector Machines*) foram aplicadas com sucesso em CA (BASILE et al., 2017; VOLLENBROEK et al., 2016; FATIMA et al., 2017). Além de outros modelos, como regressão logística (MARTINC et al., 2017; REDDY; VARDHAN; REDDY, 2017), e florestas aleatórias (MEINA et al., 2013a; MECHTI; JAOUA; BELGUITH, 2013).

Modelos mais recentes de CA passaram, entretanto, a adotar métodos baseados em redes neurais artificiais (RNA). De especial interesse para o presente trabalho, observamos que métodos de aprendizado profundo têm tido bons resultados no campo de PLN, dado o alto desempenho obtido por tais métodos e a eficiente capacidade de extração de características (GOPINATHAN; BERG, 2017; SIERRA et al., 2017). Nas seções seguintes são apresentados os conceitos básicos sobre as redes neurais do tipo *Feed-forward neural network* (FFNN), uma das arquiteturas mais tradicionais de RNAs, e duas arquiteturas

especializadas, chamadas de redes neurais de convolução (CNN) (do inglês, *Convolutional Neural Network*) e redes neurais recorrentes (RNN) (do inglês, *Recurrent Neural Networks*).

2.3.1 Redes neurais artificiais (FFNN)

As RNAs são métdos de aprendizado de máquina baseados no sistema biológico nervoso (GOODFELLOW; BENGIO; COURVILLE, 2016) e utilizam-se dos conceitos de neurônios como unidades de processamento, sinapses como canal de conexão entre estas unidades e impulsos elétricos como forma de ativação de cada unidade. O modelo de *Feed Forward Neural Network* (FFNN) foi o primeiro tipo de rede neural desenvolvido. Sua variação mais básica é conhecida como *Perceptron*, introduzida em (ROSENBLATT, 1958). Métodos de RNAs têm o objetivo de aproximar funções f^*. Por exemplo, para uma função de classificação, $y = f^*(x)$, dados exemplos de entrada x, o método tem o objetivo de mapear cada valor de entrada para um valor de saída y. A rede define um mapeamento $y = f(x; \theta)$ e aprende o valor do conjunto de parâmetros θ (GOODFELLOW; BENGIO; COURVILLE, 2016).

De modo geral, as redes do tipo FFNNs apresentam fluxos de informações propagados para frente (*Forward Propagation*) através da rede. As informações são processadas a partir dos exemplos de entrada x, pelos cálculos da cadeia de funções $f*$, e finalmente, para a saída y (NIELSEN, 2018). Existem variações deste método, sendo a principal o algoritmo de *Back Propagation* (NIELSEN, 2018).

Estes métodos são chamados de redes por causa de sua representação compondo diferentes funções combinadas. Por exemplo, dada três funções, $f1, f2, f3$, estas são conectadas em uma cadeia para formar $f(x) = f_1(f_2(f_3(x)))$. Tal estrutura é a mais comumente usada em redes neurais. Neste caso, f_1 é a primeira camada de processamento da rede, f_2 a segunda camada, e assim sucessivamente. O tamanho total da cadeia representa a profundidade da rede. (GOODFELLOW; BENGIO; COURVILLE, 2016). As arquiteturas de RNAs podem variar de complexidade a medida que sua estrutura é projetada para mais unidades de neurônios e mais camadas de grupos de unidades de neurônios. Diversos tipos de arquiteturas foram desenvolvidas para desempenhar tarefas específicas, como o caso de CNNs (GOODFELLOW; BENGIO; COURVILLE, 2016), inicialmente para visão computacional. Algumas delas são apresentadas nas seções seguintes.

2.3.2 Redes neurais recorrentes (RNNs)

O conceito de RNN foi introduzido em Rumelhart, Hinton e Williams (1986) e apresenta redes neurais para processamento de dados sequênciais, i.e., sentenças. De modo geral, as RNNs possuem um comportamento variável no tempo e têm a capacidade de abstrair características a partir da ordem das palavras, sílabas, caracteres, fonemas etc. Este tipo de comportamento é especialmente útil em dados que possuem características sequenciais. Tais dados são frequentes em problemas como reconhecimento de fala e PLN em geral (NIELSEN, 2018).

A arquitetura tradicional de uma RNN é composta por células de um único neurônio, que processam os dados de entrada x sequencialmente, permitindo a persistência da informação h a cada iteração t. Entretanto, o estudo de Bengio, Simard e Frasconi (1994) sugere que RNNs tradicionais apresentam dificuldades em processar informações com longas sequências de dados. Este caso é muito comum em tarefas de PLN quando é necessário extrair informações de documentos longos, o que leva a necessidade do aumento da janela de informação t armazenada pela rede.

Como solução para este problema, as redes de memória de curto e longo prazo (do inglês, *Long Short Term Memory* - LSTM) foram introduzidas em Hochreiter e Schmidhuber (1997) e apresentam melhorias no modelo padrão de RNN. As LSTMs são especializadas em processar sequência de dados com dependências de longo prazo, e apresentam combinações de portas (do inglês, *gates*) de esquecimento f (do inglês, *forget*), entrada i (do inglês, *input*) e saída o (do inglês, *output*).

A arquitetura da LSTM é composta por três portas e uma célula de estado de memória. Cada célula na LSTM pode ser calculada de acordo com as equações 2 à 7, obtidas de Hochreiter e Schmidhuber (1997).

$$i_t = \sigma(x_t U^i + h_{t-1} W^i) \tag{2}$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \tag{3}$$

$$o_t = \sigma(x_t U^o + h_{t-1} W^o) \tag{4}$$

$$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^g) \tag{5}$$

$$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t) \tag{6}$$

$$h_t = \tanh(C_t) * o_t \tag{7}$$

Cada célula é composta por uma porta de *input* i, *forget* f e *output* o. A matriz de pesos W conecta o estado anterior (h_{t-1}) com o atual e a matriz de pesos U conecta as entradas (x_t) com o estado atual. A porta de entrada define quanto de informação do estado calculado para a entrada atual a rede deve processar. A porta de esquecimento define quanto de informação do estado anterior a rede deve processar. E por fim, a porta de saida define quanto de informação do estado atual a rede deve explor para a próxima etapa de aprendizado.

C é a combinação do estado anterior C_{t-1} multiplicado pela porta de esquecimento e o estado atual $\tilde{C}$ multiplicado pela porta de entrada. Este processo permite que a rede ajuste o quanto de informação antiga f e quanto de informação nova i deve manter em memória. O estado oculto de saída h_t é calculado multiplicando a memória com a porta de saída.

O mecanismo de *self-attention* (WANG et al., 2016; BAHDANAU; CHO; BENGIO, 2015) fornece um conjunto de vetores de pesos com base nos estados ocultos h da LSTM, e permite extrair uma representação ponderada das partes mais importantes dos dados de entrada. As equações 8 à 10 ilustram as operações de *self-attention*, obtidas de (BAHDANAU; CHO; BENGIO, 2015).

$$u_{it} = tanh(W_w * h_{it} + b_w) \tag{8}$$

$$\alpha_{it} = \frac{exp(u_{it}^T * u_w)}{\sum exp(u_{it}^T * u_w)} \tag{9}$$

$$\tilde{h_i} = \sum_t \alpha_{it} * h_{it} \tag{10}$$

O mecanismo de atenção é composto por um score de atenção u, distribuição propabilística de atenção α e saída da atenção s. Este último, é utilizado para a camada de saída de classificação da rede, usando uma função de *softmax*.

O score de atenção u_{ij} é obtido através de uma MLP de uma única camada com os estados ocultos h_{it} da LSTM, e então mensurada a importância de cada informação como similaridade de u_{ij} com o vetor de contexto u_w para obter a distribuição de atenção normalizada α_{it}, através de uma função de *softmax*. Após isso, é calculado o vetor de saída $\tilde{h_i}$ como uma soma ponderada dos estados ocultos h_{it} e seus pesos α_{it}. O vetor de contexto

pode ser interpretado como uma representação de alto nível de quais informações a rede deve prestar mais atenção, sobre todas as informações processadas na memória da rede. O vetor de contexto u_w é inicializado aleatoriamente e aprendido em conjunto durante o processo de treinamento da rede.

Alguns estudos da área de PLN apresentam resultados significativos com o uso de LSTMs, em especial para as tarefas de CA (KIM et al., 2017; GOPINATHAN; BERG, 2017). Mais recentemente, redes do tipo *Gated Recorrent Unit* (GRU) foram introduzidas em Cho et al. (2014), e são essencialmente uma variação simplificada do modelo de LSTM. A proposta desta rede é produzir resultados similares às redes RNNs já existentes, mas com maior eficiência no tempo de execução e generalização. Os resultados apresentados por Chung et al. (2014) são competitivos com os das redes LSTMs.

2.3.3 Redes neurais convolutivas (CNNs)

O conceito de CNN foi desenvolvido em LeCun et al. (1989) e difere estruturalmente e computacionalmente de variações tradicionais de redes neurais (NIELSEN, 2018). É inspirado biologicamente em como o córtex visual de animais é organizado, e como estes respondem a estímulos em sobreposições de áreas no campo visual, conhecido como campos receptivos. Inicialmente, CNNs foram empregadas em tarefas de identificação de caracteres numéricos (i.e., classificação de imagens) (LECUN et al., 1989). Devido aos resultados promissores no campo de visão computacional, estudos recentes de PLN têm explorado o uso de CNNs em diversas tarefas (ZHANG; ZHAO; LECUN, 2015a; KIM, 2014a; LAI et al., 2015). No campo de PLN, as CNNs apresentam capacidade de extrair características de expressões e frases a partir de combinações de palavras, sílabas ou caracteres. .

A arquitetura típica de uma CNN (KIM, 2014b) para PLN é composta por componentes de convolução 1D (i.e., de uma dimensão) e agrupamento máximo. As operações de convolução c_i envolvem um *filtro* de tamanho $w \in \mathbf{R}$, que é aplicado a um intervalo de palavras h para produzir uma nova representação dos dados de entrada. As equações 11 12 e 13 apresentam, de forma resumida, as operações destes componentes, de acordo com Kim (2014b).

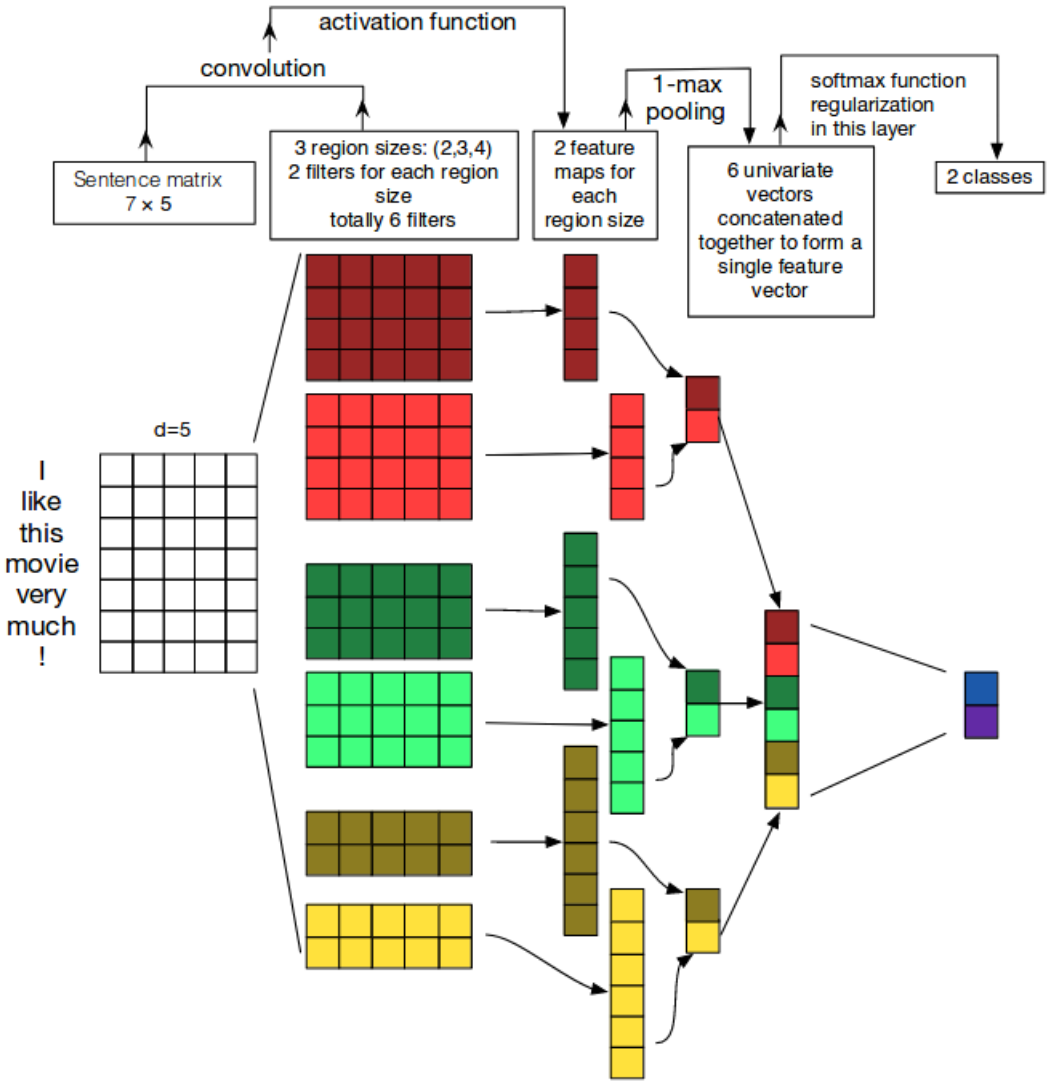
$$c_i = f(w.x_{i:i+h-1} + b) \tag{11}$$

$$\vec{c} = [c_1, c_2, ..., c_{n-h+1}] \tag{12}$$

$$m = max(\vec{c}) \tag{13}$$

O filtro é aplicado para cada possível janela nos dados de entrada (equação 11), para então produzir um mapeamento de características (equação 12). A operação de agrupamento máximo (equação 13) é aplicada sobre o mapeamento de características, obtendo o valor máximo correspondente a um filtro particular. O objetivo é capturar as informações mais importantes, com os maiores valores, para cada mapeamento.

Figura 2 – Arquitetura típica de CNN para PLN



Fonte: (ZHANG; WALLACE, 2017)

A figura 2 ilustra uma arquitetura típica de CNN para PLN. Os dados de entrada, no exemplo, a sentença: "I like this movie very much!", é representado por um vetor de tamanho 7 e dimensão 5. Neste vetor, são aplicados dois filtros de tamanhos 2,3 e

4. O resultado da operação de convolução é concatenado em um vetor de mapeamento de características, formando 2 vetores para cada tamanho de filtro. Para cada vetor de mapeamento, é aplicada a operação de agrupamento, do tipo agrupamento máximo (do inglês, *max pooling*), obtendo o valor máximo de cada vetor. Estes valores são concatenados e formam um vetor de características. Ao final, as informações são processadas por uma função de ativação *softmax*, resultando na distribuição probabilistica da classificação final.

3 Revisão Bibliográfica

Este capítulo apresenta uma revisão bibliográfica exploratória sobre estudos de CA. Os estudos aqui considerados foram selecionados a partir dos repositórios ACL Anthology, Scopus, IEEE e PAN-CLEF, e foram privilegiados artigos mais recentes e de maior fator de impacto. Para facilitar a discussão, esta revisão é dividida em estudos que usam abordagens tradicionais de aprendizado de máquina (Seção 3.1) e estudos baseados em métodos de aprendizado profundo (Seção 3.2). Além disso, é apresentado um resumo dos estudos e considerações finais (Seção 3.3) sobre esta revisão.

3.1 Abordagens tradicionais de aprendizado de máquina

Tradicionalmente, modelos de CA fazem uso de métodos de regressão logística, árvores aleatória e máquina de vetores de suporte (SVM), combinados a métodos de representações de textos baseados em contagem de palavras (BOW), frequência (TF), sequência (n-gramas) e cômputo de características específicas para cada tipo de domínio e idioma. Uma coleção de estudos recentes desta natureza é descrita a seguir.

3.1.1 Modelo de caracterização de gênero e faixa etária usando atributos de segunda ordem (SOA)

O estudo em López-Monroy et al. (2014) trata do problema de caracterização de gênero e faixa etária usando um modelo baseado em relacionamento entre documentos pertencentes a uma mesma classe com o classificador LibLinear (HALL et al., 2009). Os conjuntos de dados são do idioma inglês, espanhol e português. O experimento apresenta desempenho superior ao vencedor da competição PAN-CLEF 2013 (RANGEL et al., 2013) e foi o melhor participante da competição PAN-CLEF 2014 (RANGEL et al., 2014).

Os conjuntos de dados utilizados no estudo são da competição PAN-CLEF 2014, contendo as informações de gênero e faixa etária. O modelo propõe o uso de atributos de segunda ordem (SOA) para geração de características que representam relacionamentos entre documentos e subperfis de autores. Para isso, consideram-se três técnicas: representação de termos em um espaço de perfis, representação de documentos em um espaço de perfis e a geração de subperfis. São criados novos atributos de subperfis para faixa etária e gênero,

e então são utilizados dois classificadores para comparar e avaliar o modelo proposto. São usados como características os 3.000 termos mais frequentes, processados com o algoritmo LibLinear sem nenhum tipo de otimização de parâmetros. Segundo o estudo, considerar o relacionamento entre documentos ajudou a alcançar uma melhor discriminação entre vários perfis de autores, e com atributos gerados automaticamente o classificador manteve boas taxas de acurácia para dados desbalanceados.

3.1.2 Modelo de caracterização de gênero e faixa etária usando análise semântica latente (LSA)

O estudo em Meina et al. (2013a) apresenta um modelo para as tarefas de caracterização de gênero e faixa etária usando uma variedade de características estatísticas extraídas de conjuntos de dados nos idiomas inglês e espanhol. A avaliação foi realizada considerando o sistema de *baseline* da competição PAN-CLEF 2013 (RANGEL et al., 2013) e apresentou o melhor resultado para gênero e o segundo melhor para faixa etária nesta competição.

As características extraídas incluem número de frases, palavras e paragráfos, nível de legibilidade, análise de palavras de emoção, *emoticons* e tópicos estatísticos derivados da análise semântica latente (LSA). Foram utilizados os conjuntos de dados disponibilizados pela competição PAN-CLEF 2013, que abrange autores de publicações do Twitter. Foram extraídas 476 características para gênero e 311 para faixa etária, desconsiderando-se publicações identificadas como *spam*. A avaliação foi realizada levando-se em consideração o sistema de *baseline* da competição e outros sistemas competidores. O experimento usou o classificador Random Forest (HALL et al., 2009) e apresentou os melhores resultados da competição, porém o modelo obteve um dos piores desempenhos em tempo de processamento.

3.1.3 Modelo de caracterização de gênero, faixa etária e personalidade usando combinação de atributos de segunda ordem (SOA) e análise semântica latente (LSA)

Em Álvarez-Carmona et al. (2015) é apresentado um modelo baseado na combinação de representações discriminativas e descritivas para as tarefas de caracterização de gênero, faixa etária e traços da personalidade nos idiomas inglês, espanhol, italiano e holandês. Os

conjuntos de dados são da competição PAN-CLEF 2015 (RANGEL et al., 2015). O estudo apresenta resultados significativos com a combinação de atributos de segunda ordem (SOA) e análise semântica latente (LSA).

O estudo explora o uso de SOA e técnicas de LSA para destacar propriedades descriminativas e descritivas, respectivamente. O experimento consiste na extração das propriedades LSA e SOA, avaliação das características de forma separada e combinada, e por fim uma comparação com a técnica de representação de palavras BOW usando os algoritmos SVM e LibLinear (HALL et al., 2009).

O estudo concluí que a tarefa de caracterização de gênero, quando considerada representações individuais de LSA, obtém bons resultados e ultrapassa os resultados apresentados por BoW em todas os idiomas, e que o resultado de LSA e SOA combinados são melhores apenas no idioma inglês. Para a tarefa de caracterização de faixa etária, a representação LSA obtém melhores resultados, mas a combinação de LSA e SOA obtém uma melhora nos conjuntos de dados do idioma inglês e espanhol. A combinação de propriedades discriminativas e descritivas melhoraram os resultados significativamente, além de apresentar desempenho superior a técnica de BOW. O estudo apresenta desempenho superior aos demais trabalhos da competição PAN-CLEF 2015.

3.1.4 Modelo de caracterização de gênero e variação de idioma usando n-gramas de palavras e de caracteres

Em Basile et al. (2017) é apresentado o modelo N-GrAM, baseado em n-gramas de palavras e caracteres, para caracterização de gênero e variação de idioma em inglês, espanhol, português e árabe. O estudo avalia os algoritmos Decision Tree, MLP, Naive Bayes e SVM. O algoritmo SVM obtém melhores resultados entre os sistemas competidores da PAN-CLEF 2017 (RANGEL et al., 2017), o uso de n-gramas de palavras e caracteres provam ser modelos mais robustos para caracterização de gênero e variação de idioma.

O estudo usa os conjuntos de dados disponibilizados pela competição PAN-CLEF 2017, seguindo a distribuição em quatro idiomas e suas variações de idioma baseada em localidades. O idioma inglês possui as variações Australia, Canadá, Grã-Bretanha, Irlanda, Nova Zelândia e Estados Unidos. O idioma espanhol possui as variações Argentina, Chile, Colombia, México, Peru, Espanha e Venezuela. O idioma português possui as variações Brasil e Portugual, e o idioma árabe possui as variações Egito, Golfo, Levante, Magreb.

Considera-se nesta tarefa um modelo de classificação que possa capturar as características gerais em dois eixos, gênero e variação de idioma, e em quatro idiomas. Para resolver este problema, o estudo avalia o uso de conjuntos adicionais para analisar o desempenho da tarefa de caracterização de gênero como os conjuntos de dados disponibilizados pela competição PAN-CLEF 2016 (RANGEL et al., 2016), porém não apresenta bom desempenho. Outra alternativa considerada foi o uso do conjunto de dados Twitter 14k, usado para calcular a relação entre o uso de palavras e gênero do autor. Este conjunto demonstrou-se ser um *baseline* razoável para a tarefa de caracterização de gênero. O estudo também avalia o uso da técnica de etiquetas de POS, porém sem desempenho satisfatório, e técnicas de tokenização e extração de características como *emoticons* e exclusão de padrões específicos de palavras, como publicações iniciadas com palavras maiúsculas, minusculas etc. Por fim, avalia ainda os textos de autores com nomes de lugares, porém não apresentam resultados significativos. É utilizado o modelo de Vollenbroek et al. (2016). O estudo tenta criar um modelo único de caracterização para gênero e variação de idioma, porém não apresenta resultado satisfatório.

Para avaliação, o estudo analisa diferentes abordagens, como a implementação do algoritmo *FastText* (JOULIN et al., 2016) de representação de palavras e classificação de sentenças em conjunto com a técnica de tokenização da biblioteca *NLTK Tokenizer* (LOPER; BIRD, 2002). Visto que o conjunto de dados utilizado é pequeno e possui poucos dados de treinamento, essa última abordagem apresenta desempenho inferior quando comparado ao modelo de n-gramas e SVM. O classificador SVM apresenta o melhor desempenho, usando a implementação *LinearSVM* da biblioteca *Scikit-learn*, que usa n-gramas de 3 e 5 caracteres e n-gramas de 1 e 2 palavras com ponderação TF-IDF. O resultado é comparado com o sistema de *baseline* LDR baseado em representação de baixa dimensionalidade. Os dados adicionais de treinamento, características ditas "inteligentes" e características extraídas manualmente pioram o desempenho do modelo.

O estudo sugere que características criadas manualmente servem apenas para dificultar a capacidade do algoritmo de aprendizado de máquina de encontrar padrão em um conjunto, e sugere que seja melhor concentrar esforços na otimização de parâmetros em vez de engenharia de recursos. O modelo proposto apresentou o melhor resultado entre os competidores da PAN-CLEF 2017, obtendo o desempenho de 0,82 de acurácia para a tarefa de caracterização de gênero e 0,91 para variação de idioma. O modelo apresentou desempenho inferior para a tarefa de variação de idioma apenas para o sistema de *baseline*.

O estudo defende que o algoritmo SVM é a melhor escolha para conjuntos de dados de tamanho similar ao utilizado na competição PAN-CLEF 2017. Para conjuntos de dados de tamanho maior, é sugerido que abordagens baseadas em redes neurais podem obter melhores resultados.

3.1.5 Modelo de caracterização de gênero, faixa etária e personalidade usando combinação de *n*-gramas de caracteres e de POS

Em González-Gallardo et al. (2015) é apresentado uma modelo de *n*-gramas de caracteres e *n*-gramas de POS para caracterização de gênero, faixa etária e traços da personalidade em conjuntos nos idiomas inglês, espanhol, italiano e holandês. O estudo analisa diversas técnicas de extração de características e as avalia usando o algoritmo SVM com kernel linear LinearSVC (HALL et al., 2009). O modelo apresenta o segundo melhor resultado e o melhor tempo de processamento na competição PAN-CLEF 2015 (RANGEL et al., 2015).

Consideram-se características estilísticas representadas por *n*-gramas de caracteres e *n*-gramas de POS para classificar as publicações de usuários do Twitter. Com o objetivo de extrair o maior número possível de características codificadas em uma publicação, tais como *emoticons*, caracteres maiúsculos e caracteres repetidos, o experimento faz uso da ferramenta *FreeLing*[1] de etiquetas de POS para análise de informações morfológicas.

O experimento foi conduzido seguindo as atividades de rotulação de publicações a partir de etiquetas de POS, extração de informações de caracteres e criação de vetores de características para ser usado no treinamento do modelo. Para as tarefas de caracterização de gênero e faixa etária, é considerada a implementação de um classificador, enquanto que para a tarefa de caracterização de traços da personalidade, dado a natureza contínua de valores, foi incialmente considerado um problema de regressão. Após testes com o conjunto de dados, entretanto, foi observado que o modelo apresentou baixo desempenho em alguns algoritmos de regressão, e assim, foi implementado um algoritmo de classificação.

A avaliação é separada em duas etapas: para as tarefas de gênero e faixa etária foi usada a medida de acurácia, e para traços de personalidade a medida RMSE. Com a informação de *n*-gramas de caracteres foi possível extrair *emoticons*, identificar o uso exagerado de pontuações e caracteres, e informações emocionais codificadas nas publicações.

[1] http://nlp.lsi.upc.edu/freeling/

Já com a informação de POS n-gramas nos idiomas inglês e espanhol foi possível capturar séries mais representativas de duas e três características gramaticais. Nos idiomas restantes foi possível capturar as características gramaticais mais frequentes. O estudo apresenta um modelo com desempenho considerado satisfatório, e conclui que o uso de n-gramas de caracteres e de POS são boas opções para textos densos por causa da sua capacidade de extração de informação.

3.1.6 Modelo de caracterização de gênero usando combinação de n-gramas de POS e frequênca de termos (TF-IDF)

Em Reddy, Vardhan e Reddy (2017) é proposto um modelo de representação de documentos para caracterização de gênero no idioma inglês usando ponderação de documentos e termos com combinações de n-gramas de POS e frequência de termos (TF-IDF). O modelo é avaliado com sistemas tradicionais de representação de textos e apresenta desempenho superior quando comparado a um modelo BOW.

O modelo faz uso de conjunto de dados composto de informações de gênero de autores de avaliações de hotéis do site TripAdvisor. No total, são coletados 4.000 avaliações no idioma inglês, sendo 2.000 para gênero masculino. O estudo realiza comparações entre os modelos BOW usando n-gramas, representação ponderada de n-gramas de POS e modelo de ponderação de documentos com os 1.000 termos mais frequentes usando n-gramas de POS.

A medida de acurácia é empregada para avaliação, e são aplicados os algoritmos *Naive Bayes Multinomial* (Probabilístico), *Simple Logistic* (Funcional), IBK (Lazy), *Bagging* (Ensemble/meta) e *RandomForest* (Árvore de decisão). Todos os algoritmos foram avaliados usando validação cruzada *k-fold* ($k = 10$). Os algoritmos *Naive Bayes Multinomial* e *Simple Logistic* obtém resultados satisfatórios para caracterização de gênero. O uso de n-gramas de POS não apresentou resultado suficiente para melhorar o desempenho do modelo de caracterização de gênero.

3.1.7 Modelo de caracterização de gênero e variação de idioma usando combinação de n-gramas e frequência de termos (TF-IDF)

O estudo em Martinc et al. (2017) apresenta um método para pré-processamento de publicações do Twitter, extração de características, características de ponderação e modelos de classificação para as tarefas de caracterização de gênero e variação de idioma a partir dos conjuntos de dados da competição PAN-CLEF 2017 (RANGEL et al., 2017). O modelo proposto é avaliado com sistemas de *baseline* e apresenta os melhores resultados para o reconhecimento de variações no idioma português (i.e., português brasileiro e europeu) na competição.

O modelo proposto é um classificador baseado em regressão logística, no qual as características principais são diferentes tipos de n-gramas de caracteres e palavras. As características adicionais incluem n-gramas de POS, *emoticons*, análise de sentimentos e listas de palavras de variedades de idiomas. O modelo é avaliado pelo sistemas de baseline da competição PAN-CLEF 2017 e obtém os melhores resultados no conjunto de testes no idioma português em termos de gênero e variação de idioma. O pior desempenho foi obtido no conjunto de testes no idioma árabe.

As características usadas pelo modelo são unigramas e bigramas de palavras, tetragramas de caracteres, trigramas de pontuações e tetragramas de sufixo de caracteres. Todas as características usam ponderação TF-IDF e os parâmetros 10% e 80% para frequência mínima e máxima, respectivamente. São analisadas as características de trigramas de POS, contagem de *emoticons*, análise de sentimento e lista de palavras específicas para variação de idioma. Modelos de representação de palavras usando as técnicas de modelagem de tópicos *TruncatedSVD*[2] e *Word2Vec embeddings* (MIKOLOV et al., 2013b), além de características de contagem de palavras, contagem de pontuações e características estatísticas como tamanho do documento e média de tamanho de palavras também foram avaliadas usando validação cruzada *k-fold* ($k = 10$), porém essas técnicas não melhoraram o desempenho e não foram consideradas no modelo final.

O estudo avalia quatro algoritmos para definir o modelo de classificador vencedor: SVM, regressão logística, árvores aleatórias e *XGBBoost* (do inglês, Extreme Gradient Boosting), além da combinação dos classificadores de regressão logística e votação ma-

[2] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

joritária entre os algoritmos de regressão logística, Linear SVM e árvores aleatórias. O melhor resultado foi obtido com o algoritmo de regressão logística.

Foi considerado definir pesos às características para cada tarefa e idioma separadamente, porém isso não melhorou o resultado do modelo. Uma configuração única de pesos foi assim considerada, e provou-se ótima para ambas as tarefas e para todos os idiomas.

O modelo final obteve o melhor resultado para o conjunto de dados no idioma português, alcançou 0,86 e 0,98 de acurácia, respectivamente, para as tarefas de caracterização de gênero e variação de idioma. Para a tarefa de gênero, o pior resultado foi obtido no idioma espanhol, enquanto o idioma árabe apresentou o pior desempenho para a variação de idioma.

O estudo apresenta o segundo melhor resultado entre os sistemas competidores da PAN-CLEF 2017. O modelo baseado em n-gramas confirma os resultados relatados em trabalhos anteriores e aponta n-gramas de carateres como o modelo mais bem sucedido para as tarefas de CA.

3.1.8 Modelo de caracterização de gênero e faixa etária usando n-gramas e etiquetas de POS independentes de domínio

Em Vollenbroek et al. (2016) é apresentado um modelo linear SVM para caracterização de gênero e faixa etária usando características estilísticas independentes de conjuntos de dados nos idiomas inglês, espanhol e holandês. Foram definidas diversas características a fim de observar o seu desempenho em conjuntos de dados de diversos tipos de texto. O modelo final foi avaliado com o sistema de *baseline* da competição PAN-CLEF 2016 (RANGEL et al., 2016) e sistemas de competidores.

Os conjuntos de dados usados correspondem às coleções da competição PAN-CLEF 2016, que foram divididas em dados de treinamento a partir de conjuntos de publicações do Twitter e dados de testes a partir de conjuntos mistos de outras redes sociais. O estudo tem o objetivo de extrair características estilísticas para avaliar a portabilidade do modelo para as tarefas de caracterização de gênero e faixa etária em textos de outros tipos de mídia.

As tarefas de caracterização de gênero e faixa etária foram consideradas como problemas de classificação binária e multiclasse, respectivamente. As características selecionadas para o experimento são n-gramas de palavras, palavras iniciais em maiúsculo,

palavras completas em maiúsculo, palavras com pontuação ao final, número de pontuação por sentença, média do tamanho de palavras, média do tamanho de sentenças, número de gírias, nível de riqueza de vocabulário, palavras específicas para categoria de autor, etiquetas de POS, *emoticons*, atributos de segunda ordem (SOA) e outras. O critério de seleção das características foi a acurácia obtida a partir do conjunto de dados do Twitter e as características com menor acurácia foram selecionadas para evitar especialização quando avaliadas com outros tipos de conjuntos de dados. Todas as características foram testadas usando o modelo SVM e validação cruzada *k-Fold* ($k = 5$). As características de *emoticons* apresentam melhores resultados para o idioma espanhol do que nos demais idiomas. Além disso, o uso de contagens de n-gramas se mostrou crucial para as tarefas de CA em gêneros textuais mistos. Nenhuma das características separadas teve efeitos consideráveis para o desempenho do modelo. Por este motivo, foram usadas todas as características para testar o modelo final. O modelo proposto alcançou o melhor resultado para caracterização de faixa etária, enquanto que a caracterização de gênero alcançou o sétimo melhor desempenho na competição.

3.1.9 Modelo de caracterização de gênero e faixa etária usando n-gramas de palavras e de caracteres

O estudo em Fatima et al. (2017) compara modelos multilíngue e de tradução para tratar do problema de CA de gênero e faixa etária usando os algoritmos J48, Random Forest, SVM e Naive Bayes (HALL et al., 2009). São consideradas as características estilística e de conteúdo. Os resultados apresentam as informações de n-gramas como as mais discriminativas para caracterização de gênero e faixa etária.

São considerados conjuntos multilíngue RUEN-AP-2017 (FATIMA et al., 2017), contendo textos nos idiomas Roman Urdu e inglês, e conjuntos de textos traduzidos de RUEN-AP-2017 no idioma Roman Urdu para o inglês. O conjunto de dados RUEN-AP-2017 contêm textos de 479 usuários do Facebook, com uma média de 2.156 palavras por usuário e total de 1.039.889 palavras. Os textos contêm 479 rótulos de gênero e 479 de faixa etária.

As características estilísticas consideradas objetivam capturar o estilo de escrita de autores e a extração de métricas estatísticas sobre textos baseadas em léxico de palavras e caracteres. Estas características são a média de tamanho de sentenças, média de tamanho

de palavras, número de palavras, pontuações e outras. Além destas, é considerada a medida de riqueza de vocabulário usando o raio $\frac{N}{V}$, onde V é o tamanho do vocabulário e N é o número de palavras do texto.

Os modelos baseados em características de conteúdo consideram a combinação de caracteres para prover informações úteis sobre o conteúdo do texto. Este estudo aplica modelos de n-gramas de palavras com variação de n entre 1 e 3, e n-gramas de caracteres com variação de n entre 2 e 10. O estudo também considera três métricas de seleção de características: *Information Gain* (IG), *Gain Ratio* (GR) e *Chi Square* (Chi).

Para o experimento, foram considerados os quatro algoritmos citados, e 64 características estilísticas e de conteúdo. A avaliação foi realizada usando validação cruzada *k-fold* ($k = 10$) e medidas de acurácia e área sob a curva ROC (AUC). Além disso, foi explorado o impacto da tarefa de tradução no resultado final dos modelos, e foi criado manualmente um dicionário bilíngue para este estudo, usado para tradução de textos de Roman Urdu para inglês.

O estudo apresenta os melhores resultados usando características de conteúdo com n-gramas de palavras e caracteres usando o algoritmo Random Forest. O modelo multilíngue usando características estilísticas apresenta acurácia de 0,60 para faixa etária e 0,70 para gênero. O algoritmo Random Forest apresenta o melhor resultado para faixa etária e o algoritmo Naive Bayes para gênero. O modelo multilíngue usando características de conteúdo apresenta acurácia de 0,72 para faixa etária usando 6-gramas de caracteres e acurácia de 0,87 para gênero usando unigramas de palavras e trigramas de caracteres. Neste caso, o algoritmo Random Forest apresenta os melhores resultados para faixa etária e gênero.

O modelo usando conjunto de dados traduzido e características baseadas em conteúdo obteve desempenho satisfatório. Para a tarefa de gênero, o modelo apresentou acurácia de 0,81 usando o classificador SVM e unigramas de palavras e trigramas e 8-gramas de caracteres. Para a tarefa de faixa etária apresentou acurácia de 0,75 usando Naive Bayes e bigramas de palavras.

As informações de n-gramas de caracteres apresentam as características mais discriminativas para as tarefas de caracterização de faixa etária e gênero, e todos os testes usando informações de n-gramas ultrapassam o sistema de *baseline*. A tarefa de tradução não obteve bom desempenho e apresenta resultado inferior aos modelos multilíngue para gênero.

3.1.10 Modelo de caracterização de gênero e faixa etária usando técnicas de frequência de termos

O estudo em Mechti, Jaoua e Belguith (2013) propõe um modelo para a tarefa de caracterização de gênero e faixa etária baseado nos 200 termos mais frequentes por perfil de autor usando o classificador de árvore de decisão J48 (HALL et al., 2009). O modelo apresenta a segunda melhor acurácia na competição PAN-CLEF 2013 (RANGEL et al., 2013).

Os termos considerados no estudo foram divididos por tipos de características, tais como determinantes, preposições, pronomes, palavras relacionadas ao amor, palavras comuns usadas por adolescentes etc. Ao todo, são considerados 25 características para os dados de treinamento em inglês. As características foram usadas para treinar um classificador de árvore de decisão que apresentou desempenho inferior aos demais sistemas competidores da PAN-CLEF 2013, sendo o mais lento. O classificador apresentou a segunda melhor acurácia para gênero na competição. No entanto, a precisão para faixa etária não apresentou desempenho similar.

3.1.11 Modelo de caracterização de faixa etária e renda usando características de estilometria e sintaxe

Em Flekova, Preotiuc-Pietro e Ungar (2016) é apresentado um estudo exploratório da relação entre características de estilometria e sintaxe com os atributos de faixa etária e renda de conjuntos de dados no idioma inglês, usando modelos de regressão linear *Elastic Net* e SVM (HALL et al., 2009).

Foram utilizadas duas grandes coleções de dados de milhares de usuários do Twitter, rotulados com os atributos de faixa etária e renda. O estudo faz a extração de cinco características: o tamanho médio das publicações em nível de palavras e caracteres, nível de legibilidade, sintaxe e estilo de escrita. Essas características são consideradas como variáveis contínuas e o processamento é realizado com uso de modelos de regressão linear com os algoritmos Elastic Net e SVM com kernel RBF. O estilo de escrita não só possui correlação significativa com a faixa etária e renda dos autores, como também possui um peso maior para a caracterização de renda a partir da faixa etária. O estudo sugere que, ao

complementar o experimento com o nível de educação dos autores, é possível que ocorram melhoras significativas na caracterização da faixa etária.

3.1.12 Modelo de caracterização de gênero e faixa etária usando características baseadas em recuperação de informação

Em Weren et al. (2014) é apresentado um estudo comparativo entre combinações de seis grupos de algoritmos de classificação e 61 características para tratar o problema de CA de gênero e faixa etária nos idiomas inglês e espanhol. O estudo propõe a avaliação das características e algoritmos observando a medida de acurácia e medida F, e sistemas de *baseline*. As características baseadas em recuperação de informação apresentaram os melhores desempenhos.

Foram extraídas e calculadas 61 características divididas em seis tipos: tamanho de textos (caracteres, palavras e sentenças), recuperação de informação (IR) (i.e., cosseno etc.), nível de legibilidade do texto, análise de sentimento, nível de correção do texto e nível de estilo (etiquetas HTML e diversidade de palavras).

As coleções de dados da competição PAN-CLEF 2013 (RANGEL et al., 2013) foram usados neste estudo, contendo publicações de blogs e informações sobre gênero e faixa etária de autores. As informações de gênero possuem apenas duas classes (masculino e feminino), enquanto que faixa etária possui três (10s, 20s, 30s). Foram aplicadas técnicas de pré-processamento para remoção de caracteres de espaço e tokenização. Não foram removidos *stopwords* e radicais para manter as características de estilo de escrita dos autores. Foi utilizado o motor de busca Zettair[3] para extração de 30 características baseadas em recuperação de informação, e foi criada uma ponderação para cada documento por grupo de gênero e faixa etária a partir da medida coseno e Okapi BM25 (MANNING; RAGHAVAN; SCHüTZE, 2008). As características de nível de legibilidade foram calculadas usando o algoritmo *Readability*[4] e as medidas *Flesch Reading Ease* e *Flesch-Kincaid Grade Level* (KINCAID et al., 1975). As características de análise de sentimento foram baseadas no conjunto léxico *NRC Emoticon* (ZHU; KIRITCHENKO; MOHAMMAD, 2014). Características de nível de correção do texto foram baseadas em dicionários[5].

[3] http://www.seg.rmit.edu.au/zettair/
[4] http://tikalon.com/blog/2012/readability.c
[5] http://extensions.openoffice.org/en/project/english-dictionaries-apache-openoffice

Os algoritmos de classificação foram determinados a partir do estudo de Rangel et al. (2017). Ao todo, são seis grupos de algoritmos: Bayes, Functions, Lazy, Meta, Misc, Rules e Trees. Muitos destes algoritmos possuem parâmetros que podem ser ajustados para melhorar seus desempenhos, mas para estes parâmetros foram aplicados os valores padrões.

As características e os algoritmos foram avaliados comparando-os com os sistemas competidores da PAN-CLEF 2013, usando as medidas de acurácia e medida F. A avaliação de características é realizada em duas configurações. A primeira considera todas as 61 características, e a segunda considera apenas um subconjunto destas. A primeira configuração não apresenta bom desempenho ou melhora significativa nos resultados, enquanto que a segunda apresenta uma redução considerável no custo computacional e tempo de processamento. As características baseadas em IR apresentam os melhores resultados. São avaliados 55 algoritmos de classificação: o algoritmo de regressão logística, do grupo *Functions*, obteve o melhor resultado para a tarefa de caracterização de faixa etária usando a primeira configuração de características. Os algoritmos avaliados usando a segunda configuração de características apresentaram resultados significativos para caracterização de faixa etária com *Decision-trees* e *Functions*, e resultados similares entre os algoritmos *ClassificationViaRegression*, *RandomSubspace*, *MultilayerPerceptron* e *SimpleCart*.

O sistema de *baseline* é superado pelo modelo proposto, tanto para a tarefa de caracterização de faixa etária como para caracterização de gênero. As características baseadas em recuperação de informação apresentam melhores desempenhos, e o estudo defende que essas características são mais discriminativas para as tarefas de caracterização de faixa etária e gênero. Características extraídas de dicionários criados para análise de sentimento não apresentam resultados satisfatórios, e o mesmo acontece para os testes com as características de nível de legibilidade.

O número de características extraídas para a tarefa de caracterização de faixa etária é superior à tarefa de caracterização de gênero, razão pela qual grande parte dos sistemas apresentaram os melhores resultados para a primeira. A acurácia melhora em 20% quando utilizada a combinação de características baseadas em IR e o classificador J48 (HALL et al., 2009).

O tempo de processamento dos modelos também é avaliado, obtendo-se maior eficiência quando usada a segunda configuração de características. Na fase de treinamento,

o algoritmo *MultilayerPerceptron* apresenta um tempo 40 vezes mais superior ao da primeira configuração de características, considerando-se a tarefa de gênero.

O estudo conclui que o conjunto de dados disponibilizado pela competição PAN-CLEF 2013 possui muitos ruídos, motivo pelo qual as publicações de blogs tendem a ser repetidas e terem conteúdos de outros autores. Analisando o conjunto de publicações, foram encontradas 11 publicações idênticas, porém oito delas marcadas como originadas de autores masculinos e outras três de autores femininos.

3.1.13 Modelo de caracterização de gênero e faixa etária usando léxico com ponderação

O estudo em Sap et al. (2014) apresenta um modelo léxico com ponderação usando coeficientes de regressão linear multivariada e modelos de classificação para a tarefa de caracterização de faixa etária e gênero no idioma inglês. Modelos derivados de conjuntos do Facebook, blogs e Twitter apresentam poder de generalização superior quando avaliada em diferentes domínios de textos.

O estudo utiliza textos de usuários do Facebook de Kosinski e Stillwell (2012) no idioma inglês como conjunto de dados principal. São mensagens de 75.394 usuários contendo 300 milhões de palavras. Também foram considerados textos adicionais do Facebook, blogs e Twitter para avaliar o modelo léxico em diferentes domínios de textos. O conjunto de dados principal retringiu-se a usuários com pelo menos 1.000 palavras escritas e faixa etária inferior a 65 anos. O conjunto de dados adicional apresenta três subconjuntos: textos estratificados do Facebook de 3.040 usuários, publicações de blogs de 15.006 usuários e publicações do Twitter de 11.000 de usuários.

A faixa etária foi considerada um valor contínuo, e foi aplicado Regressão Linear Ridge com penalização L2. Gênero foi tratado como problema de classificação binária, e foi aplicado SVM usando kernel linear com penalização L1. O estudo ainda explora diversos algoritmos, incluindo Lasso, Elastic Net Regression e SVM com penalização L2.

Foram usadas informações de unigramas de palavras para compor o modelo léxico. Também foi aplicada a implementação Happier Fun Tokenizer[6] para manipular conteúdo de redes sociais e anotação de caracteres de *emoticons* e *hashtags*.

A avaliação foi realizada usando cinco modelos léxicos: Facebook (FB_{lex}), blogs (BG_{lex}), Twitter (T_{lex}), $FB + BG_{lex}$ e $FB + BG + T_{lex}$. Os conjuntos de dados adicionais

[6] Disponibilizado por http://www.wwbp.org/data.html

foram usados para testar o modelo léxico derivado do Facebook. Os resultados para faixa etária e gênero apresentam desempenho superior ao *baseline*, porém por causa da natureza privada dos dados de Kosinski e Stillwell (2012), não foi possível comparar seus resultados com outros sistemas. Por este motivo, foi usado o conjunto de dados do Twitter. O modelo ultrapassa os resultados dos sistemas de *baseline*, exceto para Burger et al. (2011).

O segundo teste foi realizado para avaliar o poder de generalização do modelo. Foi aplicado um filtro nos dados do Facebook para selecionar números balanceados de gêneros (masculino e feminino) e faixa etárias entre 13 e 60 anos. Visto que o conteúdo e variação de estilo destes textos podem ser específicos do Facebook, o estudo usa estes textos para treinar o modelo e avaliá-lo a partir de outros domínios, como blogs e Twitter. Os testes apresentam resultados similares ao modelo criado para domínios específicos.

O terceiro teste foi realizado usando um número reduzido de mensagens por usuário (média de 205 mensagens). O estudo avalia os modelos $FB + BG_{lex}$ para faixa etária e $FB + BG + T_{lex}$ para gênero. Os testes não apresentam resultados satisfatórios, e é observado que quanto menor o número de mensagens por usuário, menor é o desempenho do modelo.

3.1.14 Modelo de caracterização de gênero usando dicionário LIWC independente de conjuntos de dados

Em Isbister, Kaati e Cohen (2017) é apresentado um modelo de características independentes de conjuntos de dados para classificação de gênero em quatro idiomas usando o dicionário LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001). O modelo apresenta resultados satisfatórios para o idioma inglês, enquanto que nos demais, sueco, francês, espanhol e russo, tem seu desempenho limitado ao volume de textos e atualizações recentes do dicionário LIWC.

Os conjuntos de dados coletados, além de conter as informações de gênero de cada autor, apresentam categorização do conteúdo do texto em vários tópicos, como política, esportes, assuntos pessoais e viagens, balanceados por gênero. Os experimentos foram aplicados usando as mesmas configurações de validação cruzada *k-fold* ($k = 5$) e 25% do conjunto de dados para testes. Também foi usado o classificador SVM e as medidas de acurácia, precisão e revocação. O estudo faz análises sobre a frequência e importância de

cada característica para gênero e idioma e lista as 10 características mais importantes para classificação de gênero em cada idioma.

Os experimentos mostram que usando algoritmos de aprendizado de máquina com as características extraídas do dicionário LIWC é possível obter uma acurácia de 0,73 à 0,79, dependendo do idioma. O estudo sugere que o desempenho significativo do modelo de caracterização de gênero no idioma inglês se deve à utilização da versão atualizada de 2015 do dicionário LIWC para inglês (PENNEBAKER; FRANCIS; BOOTH, 2001), além do número relativamente alto de palavras por autor no conjunto de dados neste idioma, que permite ao classificador aprender com maior volume de dados e atingir melhor desempenho. O estudo privilegiou a análise de características gramaticais como artigos, pronomes, primeira pessoa do singular, palavras de função, verbos e afeto, pois em trabalhos passados foi comprovado serem importantes para distinguir gêneros. Por fim, o estudo sugere como trabalho futuro investigar mais características para melhorar os resultados.

3.2 Abordagens de aprendizado profundo

Em contraste às abordagens tradicionais discutidas na seção anterior, alguns modelos mais recentes de CA passaram a adotar métodos baseados em aprendizado profundo usando redes neurais artificiais (RNA) e representação de textos usando métodos de distribuição de palavras como *embeddings* e extração de características usando camadas de abstração de RNA. Estudos desta natureza são descritos a seguir.

3.2.1 Modelo de caracterização de gênero usando rede neural convolucional recorrente com janela de contexto

Em Bartle e Zheng (2015) é apresentado um modelo de rede neural convolucional recorrente com janela de contexto (WRCNN) para as tarefas de caracterização de gênero a partir de blogs e livros da literatura usando extensões do modelo de RCNN em Lai et al. (2015). O modelo proposto obtém desempenho médio 4% superior ao sistema de *baseline*, provando que uma abordagem que usa informação de contexto de sentenças pode produzir um peso maior e beneficiar a caracterização de gênero.

A proposta é um modelo em que, além das palavras, é considerado o contexto de sentenças em um documento. O modelo identifica uma ativação mais forte em cada

sentença, e em seguida em todo o documento, ignorando a ordem das sentenças. Para minimizar a perda de entropia cruzada, foi aplicada regularização L2 nas matrizes de peso W e foi usado gradiente descendente estocástico para o treinamento da rede.

O modelo obteve melhores resultados com os hiperparâmetros $C = 50$ e ambas as camadas ocultas com 100 dimensões. O conjunto de dados utilizado é dividido em duas categorias, blogs e livros da literatura do século 19 e 20, no idioma inglês. O conjunto de dados de blogs, disponibilizados em Mukherjee e Liu (2010), é composto por 1.679 autores masculinos e 1.548 femininos, e inclui informações como *emoticons*, pontuação irregular e erros gramaticais. O conjunto de dados de livros da literatura, coletado do projeto *Gutenberg Book*[7], conta com baixa representação de autores femininos, e por este motivo o estudo realizou uma seleção de trabalhos para alcançar número similar de amostras de autores masculinos (2.359) e femininos (2.288).

O estudo realiza a comparação de três abordagens: BOW, *Average Embeding* usando uma simples rede neural de camada oculta única e *Paragraph2vec* baseado na implementação de *Word2vec* e etiquetas POS. Para todos os modelos, foi aplicado o algoritmo de classificação SVM da biblioteca LibSVM (PEDREGOSA et al., 2011).

O modelo proposto obtém 0,86 de acurácia para o conjunto de dados de blogs, e acurácia de 0,73 e 0,71 para os conjuntos de dados da literatura do século 19 e 20, respectivamente. Todos os desempenhos foram significativamente baixos para conjuntos de dados da literatura. Os conjuntos de dados de blogs são enviesados, pois amostras para autores masculinos tendem a apresentar conteúdos associadas a contextos masculinos, e o mesmo para autores femininos. Conjuntos de dados da literatura tendem, por outro lado, a apresentar contextos similares e com menor enviesamento de conteúdo.

O modelo de BOW, apresenta o pior desempenho para o conjunto de dados de blogs. Já os modelos *Average embedding* e *Paragraph2vec* são melhores que BOW por capturar mais informações contextuais e apresentar menos problemas com dados esparsos. O modelo RCNN apresenta melhor aprendizado de ordem de palavras e estrutura de sentenças quando comparado com modelos similares ao *Paragraph2vec*. O estudo conclui que o agrupamento máximo aplicado em sentenças é capaz de obter melhor acurácia quando comparado com resultados do estado da arte em Mukherjee e Liu (2010).

[7] http://www.gutenberg.org

3.2.2 Modelo de caracterização de faixa etária usando rede neural convolucional para classificação binária

O estudo em Guimarães et al. (2017) apresenta um modelo para caracterização de faixa etária de autores de textos, entre grupos de adolescentes e adultos, usando árvores de decisão, floresta aleatória, SVM, MLP e CNN. Os modelos que apresentam os melhores resultados são os baseados em CNN.

A classificação em apenas dois grupos de faixa etária é motivada pelas grandes diferenças de comportamentos destes dois grupos em ambientes digitais. No caso do grupo de adolescentes, observam-se comportamentos relacionados a peculiaridades no estilo de escrita, como uso de gírias, abreviações, etc. Já os adultos preferem falar sobre informações positivas e são propensos a compartilhar links de conteúdos que representam os assuntos dos quais estão falando. Há também uma grande diferença quanto aos tópicos pelos quais cada grupo é interessado. Os adolescentes falam mais sobre assuntos relacionados à vida pessoal, enquanto que os adultos preferem não expor assuntos pessoais.

Os conjuntos de dados empregados são originados do Twitter e foram coletados por segmentação de tópicos como responsabilidade, esportes, saúde, política, religião, trabalho e família. O estudo faz a extração de característica relacionadas ao estilo de escrita levando em consideração o uso de pontuação, ícones de *emoticons*, número de caracteres, tamanho das sentenças, gírias, uso de URLs para compartilhamento de conteúdos, número de pessoas seguidas, número de seguidores, número total de mensagens postadas, menção a usuários (*retweet*), uso de *hashtag*, contexto de assuntos e tópicos relacionados. Além disso, o estudo realiza uma coleta adicional de dados de faixa etária e gênero por meio de questionário online. O estudo defende que todos estes itens são importantes para melhorar a assertividade e a acurácia para a tarefa de caracterização de faixa etária.

O estudo propõe o uso de CNNs para melhor capturar a semântica de textos, quando comparado com as redes neurais recursivas (LAI et al., 2015). Além das CNNs, a avaliação dos modelos é realizada usando algoritmos de classificação de textos, como árvores de decisão, floresta aleatória, SVM e MLP. A avaliação do modelo de CNN é realizada usando representação distribuída de palavras obtida com *Word2vec* (MIKOLOV et al., 2013b) e aplicando os parâmetros de taxa de aprendizagem de 0,0004, *batch size* de 128, *fator decay* de 0,001 e *momentum* de 0,9. O treinamento usou 100 épocas e a função de otimização *Softmax*.

Em uma primeira etapa, foram consideradas todas as características extraídas dos conjuntos de dados, e posteriormente foram descartadas as *hashtags* e *retweets*, pois apresentaram baixa relevância para o desempenho do modelo.

Os modelos baseados em CNNs obtém os melhores resultados, com medida F média de 0,94, sendo 0,96 para grupos de adolescentes e 0,91 para grupo de adultos.

3.2.3 Modelo de caracterização de gênero e variação de idioma usando rede neural convolucional e técnicas de representação distribuída de palavras

Em Sierra et al. (2017) um modelo de CNN é apresentado para caracterização de gênero e variação de idioma a partir dos conjuntos de dados da competição PAN-CLEF 2017 (RANGEL et al., 2017), usando técnicas de representação distribuída de palavras como entrada. O estudo sugere que esta representação possui desempenho superior ao uso de entradas baseadas em caracteres, como observado em diversas configurações de arquiteturas testadas para CA.

Os conjuntos de dados utilizado consistem de textos do idioma inglês, espanhol, português e árabe. Os dados de treinamento consistem de 10.800 usuários do Twitter, sendo 3.000 documentos para o idioma inglês, 4.200 para o espanhol, 1.200 para português e 2.400 para árabe. Para cada idioma, foram treinados modelos separados de gênero e variação de idioma.

Os textos são representados na forma de *word embeddings* e usados como entrada para a camada de convolução e para a camada de agrupamento máximo. A classificação final é feita por meio da função *softmax* aplicada à representação de texto final.

São exploradas diversas arquiteturas de CNNs, e o estudo se concentra em dois tipos de hiperparâmetros: os relacionados à entrada e os relacionados à convolução. No geral, as configurações avaliadas possuem dados de entrada baseados em palavras ou caracteres, com tamanho dos dados de entrada variando entre 50, 100, 200 e 300, número de filtros de convolução entre 1.500 e 3.000 e tamanho dos filtros de convolução de (1,2,3), (2,3,4) e (4,5,6).

Para avaliação, foi gerada uma divisão estratificada de treinamento e validação para todas as combinações possíveis de gênero e variação de idioma. Dez por cento do conjunto de treinamento foi usado para validação. A avaliação foi realizada com sistemas de *baseline* e de competidores da PAN-CLEF 2017 usando a métrica de acurácia.

O modelo apresentou bom desempenho usando sequência de palavras. Entretando, os autores defendem que o uso de CNN para CA produz desafios adicionais como ajustes de hiperparâmetros e *overfitting* rápido. No estudo, estes desafios foram contornados usando regularização *dropout* e uma arquitetura com menor números de parâmetros.

O modelo proposto foi capaz de aprender padrões significativos sem o uso de características definidas manualmente. Entretanto, ainda apresenta desempenho inferior aos modelos tradicionais que usam concatenação de conteúdo e características de estilo de escrita definidas manualmente. Os autores sugerem como trabalho futuro explorar CNNs mais profundas usando regularização forte e estratégias de *Oversampling*. Como sugerido em Ortega-Mendoza et al. (2016), o uso de sequência de textos centradas em pronomes pessoais poderiam melhor o desempenho na CNN, pois a rede identificaria apenas os exemplos mais relevantes.

3.2.4 Modelo de caracterização de gênero e faixa etária usando redes neurais recursivas de grafos

O estudo apresentado em Kim et al. (2017) propõe um modelo de classificação de vértices de grafos usando redes recursivas para identificar gênero, faixa etária e tipo de usuário do Twitter no idioma inglês. O método converte um grafo em estruturas de árvores e usa estruturas de RNNs para cada árvore.

O modelo se apoia nas propriedades de topologia da rede, conteúdo do texto e rótulo de informação. É baseado em arquiteturas de redes neurais recursivas de grafos (GRNNs) e demonstra ser muito eficiente para as tarefas de classificação de textos, permitindo captura de composição sintática e semântica (SOCHER et al., 2011; QIAN et al., 2015). Os conjuntos de dados utilizados foram coletados de 1.000 usuários do Twitter e contêm dados binários de gênero (masculino e feminino) e faixa etária (faixa etária de 18-23 e 25-30), além de dados multiclasse do tipo de usuário (indivíduo, organização ou outro).

A avaliação foi realizada comparando-se o desempenho do modelo GRNN com diversos modelos existentes: o modelo léxico produzido a partir de dados do Twitter (Léxico), o modelo linear de regressão logística, o modelo semi-supervisionado baseado em grafo (Label Propagation), o modelo não supervisionado de incorporação de vértices (Text-Associated DeepWalk) e o modelo baseado em duas redes neurais de incorporação de textos, relacionamentos e rótulos de vértices de grafos (Tri-Party Deep Network Representation).

Além destes, foram configuradas seis arquiteturas de redes neurais baseadas em grafos aplicando-se combinações de *Graph Naive Recursive Neural Unit* (GNRNN) e *Graph Long Short- Term Memory Unit* (GLSTMN), com variação de três estruturas de árvores para medir o número de graus de conexão entre usuários na rede social. A medida acurácia foi empregada, e as combinações de GLSTMN obtêm os melhores desempenhos para caracterização de gênero e faixa etária. Os modelos GNRNN e GLSTMN com estrutura de árvore com grau de conexão dois obtêm o melhor desempenho para caracterização de tipo de usuário.

3.2.5 Modelo de caracterização de gênero usando combinação de ensemble de convolução e LSTM bidirecional e representação distribuída de palavras

Em Gopinathan e Berg (2017) são apresentadas três arquiteturas de redes neurais para tratar do problema de caracterização de gênero a partir de textos do Twitter. Foram adotadas representações de nível de caracteres com camadas de convolução e LSTM bidirecional, nível de palavras com LSTM bidirecional usando representação *GloVe* (PENNINGTON; SOCHER; MANNING, 2014) e nível de documentos com *feedforward* usando *BOW*. O estudo ainda explora um método de *ensemble* combinando as três arquiteturas por voto majoritário e apresenta os melhores resultados, superando o sistema de *baseline*.

Os conjuntos de dados de treinamento são compostos por coleções recentes da competição PAN[8], somando um total de 655.268 publicações do Twitter. Já os conjuntos de testes são compostos por coleções de dados da competição Kaggle[9], contendo 12.727 publicações do Twitter.

O estudo divide as atividades em pré-processamento, extração de características e classificação. A fase de pré-processamento é similar para todos os modelos: todos usam palavras minúsculas, etiquetas de marcação, *stopwords*, lematização, remoção de pontuação, remoção de *emoticons* e remoção de textos menores que dois caracteres. O estudo adota técnicas de extração de características baseada em frequência de termos de internet (como *hashtags* e URLs), frequência de *emoticons* específicos, tamanho de publicações, métricas de análise de sentimentos (HUTTO; GILBERT, 2014) e frequência de etiquetas de POS.

[8] https://pan.webis.de

[9] Kaggle é uma plataforma Crowdsourcing para competições de análise e modelagem preditiva: https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Para classificação, foram desenvolvidas três arquiteturas de modelos baseados em aprendizado profundo para processar textos em diferentes níveis de granularidade: nível de caracteres, de palavras e de documentos.

As arquiteturas baseadas em nível de palavras tratam textos como uma sequência de palavras. Neste sentido, o mais baixo nível de *tokens* que existe são as palavras. Estas arquiteturas usam informações de *word embeddings* para representação de textos e modelos pré-treinados de *GloVe*[10] de 200 dimensões, contendo 1,2 milhões de palavras a partir de 2 bilhões de publicações do Twitter.

Após a primeira camada de *word embeddings*, é adicionada uma camada bidirecional de LSTM. Cada camada contém 250 unidades de memória. Em seguida, há uma camada de concatenação de vetores contendo 500 neurônios usando regularização *droupout* regular e recorrente. A camada de saída usa a função de ativação *softmax* e o resultado é apresentado em uma distribuição descrevendo a probabilidade para cada gênero. Com base nesta especificação, diferentes topologias foram testadas.

As arquiteturas de nível de palavras limitam-se aos modelos pré-treinados e ignoram palavras que não existem no vocabulário. Como alternativa, as arquiteturas de nível de caracteres são projetadas para ultrapassar esta limitação processando textos como uma sequência de caracteres. A representação de textos neste caso é realizada usando um índice de caracteres extraído a partir do mapeamento de todos os caracteres existentes no conjunto de dados de treinamento. Adota-se uma camada bidirecional de LSTM similar às arquiteturas de nível de palavras, porém com adição de uma primeira camada de convolução com 1024 filtros e agrupamento máxima para sumarização de dois vetores por tempo.

Arquiteturas de nível de documento usam informações de n-gramas e construção de vocabulário, como itens frequentes e contagem de palavras relacionadas aos gêneros masculino e feminino, além de informações de análise de sentimento. Adota-se neste caso uma representação de contagem de palavras BOW e frequência de termos TF-IDF, usando como classificação uma rede neural *feedforward* com três camadas ocultas contendo, respectivamente, 2048, 1024 e 512 neurônios. A camada de entrada tem tamanho similar ao vocabulário, otimizado para 10000 no melhor desempenho durante os experimentos. A camada de saída contêm um neurônio para cada tipo de gênero.

[10] Stanford GloVe embeddings

Adicionalmente, um método de *ensemble* foi proposto usando diversos modelos baseados em diferentes arquiteturas, treinados usando os mesmo dados de entrada. O resultado de cada classificação é usado na caracterização final usando as métricas de agregação de função como voto majoritário, máximo e média.

O experimento foi realizado avaliando-se o desempenho de cada classificador e então seguido de otimizações para cada um deles, separadamente. Com o objetivo de explorar o poder combinado dos modelos otimizados, foi avaliado o método de *ensemble* usando voto majoritário. A avaliação é realizada primeiramente usando conjuntos de testes para cada modelo individual, e depois usando o método de *ensemble*, e é baseada nas métricas de precisão, revocação e medida F.

Os três modelos individuais apresentam resultados similares. O modelo de LSTM bidirecional usando *word embeddings* pré-treinados de *GloVe* em nível de palavras apresenta medida F de 0,60. O modelo convolucional e LSTM bidirecional em nível de caracteres apresentam medida F de 0,59, e o modelo de nível de documento com *feedforward*, métodos tradicionais de PLN, extração de características definidas manualmente e representação de textos BOW apresentam medida F de 0,58.

Diversas versões de *ensemble* foram exploradas, combinando os três modelos treinados separadamente, e foi usado o modelo que apresentou maior confiança a partir do voto majoritário. O melhor resultado foi produzido combinando-se modelos de nível de caracteres e nível de palavras com medida F de 0,61, superando o sistema de *baseline*, baseado em regressão logística, com medida F de 0,58.

O estudo sugere que arquiteturas de aprendizado profundo são capazes de aprender diferentes aspectos do texto e, se combinado modelos colaborativamente, podem produzir resultados com maior qualidade de caracterização. Os resultados também indicam que modelos baseados em características definidas manualmente são inferiores aos modelos baseados em *embeddings* que apresentam capacidade de extrair estas características e de representá-las de forma implícita.

3.3 Considerações

Este capítulo apresentou uma revisão bibliográfica acerca dos estudos de CA. O Quadro 1 apresenta uma visão geral desta revisão, contendo colunas com a citação de

cada artigo, o conjunto de dados, o idioma, a tarefa de CA, o tipo de conhecimento e o método de aprendizado de máquina (AM) utilizado pelos estudos.

Os conjuntos de dados usados pelos estudos são, em sua maioria, da competição PAN-CLEF (do inglês, *Lab on Digital Text Forensics and Stylometry on Conference and Labs of the Evaluation Forum*), entre as edições de 2013 e 2017, além de conjuntos próprios coletados do Twitter e de Blogs, conjuntos de dados públicos baseados no Facebook, assim como avaliações de hotéis, livros da literatura do século XX e conjuntos disponibilizados pela plataforma Kaggle[11].

Os idiomas considerados pelos estudos são inglês (EN), espanhol (ES), português (PT), italiano (IT), holandês (HO) e árabe (AR). Os estudos são, quase em sua totalidade, dedicados ao idioma inglês, e encontramos apenas um exclusivo para o idioma português (Guimarães et al., 2017).

As tarefas de CA consideradas pelos estudos são gênero (G), faixa etária (I), variação de idioma (V), personalidade (P) e renda (R). Encontramos soluções para as tarefas de caracterização de gênero e faixa etária em todos os estudos, enquanto que apenas um estudo propôs solução para a tarefa de caracterização de renda (FLEKOVA; PREOTIUC-PIETRO; UNGAR, 2016). Observa-se que as tarefas consideradas pelos estudos estão intrinsecamente ligadas aos conjuntos de dados utilizados e a existência da anotação destas informações.

São considerados tipos de representação baseados em contagem e frequência (TF) de palavras, caracteres, n-gramas, POS, léxico, LIWC, *word embeddings*, CNNs (SIERRA et al., 2017; GOPINATHAN; BERG, 2017) e RNNs (KIM et al., 2017). Os estudos são, de modo geral, baseados em tipos de representação dependentes de domínio e idioma, e foram encontrados apenas três estudos utilizando representação de *word embeddings* (GOPINATHAN; BERG, 2017; SIERRA et al., 2017; Guimarães et al., 2017), e dois estudos adotando redes neurais profundas (BARTLE; ZHENG, 2015; KIM, 2014a).

Os métodos de AM considerados pelos estudos são J48, Random Forest, SVM, regressão linear, regressão logística, além de métodos de aprendizado profundo usando CNN e LSTM. Encontramos estudos em número significativo usando SVM para classificação de gênero e faixa etária, e apenas dois usando regressão linear para caracterização de faixa etária (FLEKOVA; PREOTIUC-PIETRO; UNGAR, 2016) e (SAP et al., 2014). Os estudos usando aprendizado profundo, em sua maioria, utilizam CNNs, e encontramos um

[11] https://kaggle.com

apenas usando LSTM (KIM et al., 2017) e um outro usando método de *ensemble* de CNN e LSTM (GOPINATHAN; BERG, 2017).

Quadro 1 – Resumo dos trabalhos correlatos

Estudo	**Dados**	**Idioma**	**Tarefas**	**Representação**	**Método**
Weren et al. (2014)	PAN-CLEF 2013	EN, ES	G, I	TF, IR	J48
Mechti, Jaoua e Belguith (2013)	PAN-CLEF 2013	EN, ES	G, I	TF	J48
Flekova, Preotiuc-Pietro e Ungar (2016)	Twitter	EN	I, R	Sintaxe	SVM
Meina et al. (2013a)	PAN-CLEF 2013	EN, ES	G, I	TF, LSA	RandForest
Carmona et al. (2015)	PAN-CLEF 2015	EN, ES, IT, HO	G, I, P	LSA, SOA	SVM
López-Monroy et al. (2014)	PAN-CLEF 2014 e 2013	EN, ES, PT, AR	G, I	TF, SOA	LibLinear
Vollenbroek et al. (2016)	PAN-CLEF 2016	EN, ES, HO	G, I	word n-grams	SVM
Basile et al. (2017)	PAN-CLEF 2017 e 2016	EN, ES, PT, AR	I, V	word n-grams e char n-grams	SVM
Fatima et al. (2017)	RUEN-AP-2017	EN, AR	G, I	word n-grams e char n-grams	SVM
González-Gallardo et al. (2015)	PAN-CLEF 2015	EN, ES, IT, HO	G, I, P	char n-gramas e POS n-grams	SVM
Reddy, Vardhan e Reddy (2017)	TripAdvisor	EN	G	POS n-grams, TF-IDF	RegLog
Martinc et al. (2017)	PAN-CLEF 2017	EN, ES, PT, AR	G, V	POS n-grams	RegLog
Sap et al. (2014)	MyPersonality	EN	G, I	Léxico	SVM
Isbister, Kaati e Cohen (2017)	Blogs	EN, ES, FR, RU	G	LIWC	SVM
Guimarães et al. (2017)	Twitter	PT	I	*Word embeddings*	CNN
Sierra et al. (2017)	PAN-CLEF 2017	EN, ES, PT, AR	G, V	*Word embeddings*	CNN
Gopinathan e Berg (2017)	PAN[12] e Kaggle	EN	G	*Word embeddings*	CNN, LSTM
Bartle e Zheng (2015)	Blogs e Livros	EN	G	RCNN	RCNN
Kim et al. (2017)	Twitter	EN	G, I	RNN	LSTM

Fonte: Rafael Sandroni Dias (2018)

4 Estudo exploratório

Os estudos discutidos no capítulo de revisão bibliográfica (Capítulo 3) apresentam uma visão geral de como as tarefas de CA são tratadas. Há um nível de esforço na linha de extração de características, compondo conhecimento linguístico para construção de características específicas. Estas estratégias, em geral, apresentam resultados do estado-da-arte, porém em muitos casos demanda tempo e necessidade de conhecimento prévio linguístico para a modelagem desses sistemas especialistas. Também, os estudos são frequentemente aplicados, em sua maioria, para as tarefas de caracterízação de gênero e faixa etária.

Entretanto, diversos estudos apresentam resultados do estado-da-arte em problemas de classificação de documentos da área de PLN em geral, usando modelos baseados em RNAs, tais como CNNs (ZHANG; ZHAO; LECUN, 2015b; KIM, 2014b; TAKAHASHI et al., 2018), LSTM (WANG et al., 2016) e o uso de representação distribuída de palavras. Neste sentido, observa-se a oportunidade para o desenvolvimento de modelos de CA baseados em RNAs, sem a necessidade de conhecimento prévio linguístico e permitindo o desenvolvimento de modelos de CA independentes do tipo de domínio e idioma do texto. Além disso, com base nos córpus considerados, é possível abordar um maior número de tarefas de CA, e possivelmente tarefa inéditas de CA, tais como caracterização de grau de educação, grau de religiosidade, formação em TI e posição política, no idioma português brasileiro.

Para isso, considera-se três experimentos iniciais para explorar a organização dos córpus e entendimento do problema (Seção 4.1) (DIAS; PARABONI, 2018b; HSIEH; DIAS; PARABONI, 2018), a avaliação dos modelos baseados em representação distribuída de palavras (Seção 4.2) (DIAS; PARABONI, 2018a), e por fim, a avaliação de modelos de CNNs e RNNs (Seção 4.3) (DIAS; PARABONI, 2019). Nas seções seguintes, são apresentados detalhes de tais estudos e os resultados obtidos.

4.1 Caracterização autoral de usuários do Facebook brasileiro

Neste experimento, foram desenvolvidos modelos de aprendizado supervisionado para caracterização autoral de usuários do Facebook, com base no córpus b5-post (RAMOS

et al., 2018), no idioma português brasileiro, contemplando quatro tarefas de CA, são elas caracterização de gênero, faixa etária, grau de religiosidade e formação em TI. Este experimento teve como objetivo a organização do córpus b5-post e a exploração de abordagens de aprendizado supervisionado e técnicas de representação textual, para o desenvolvimento de modelos computacionais iniciais, para esta dissertação.

O córpus b5-post contém textos de 1019 autores (i.e., usuários) do Facebook, totalizando cerca de 2,2 milhões de palavras. O córpus é rotulado com informações de gênero, faixa etária, grau de religiosidade e formação em TI (i.e., mais detalhes sobre as tarefas podem ser encontradas no capítulo de organização dos córpus).

4.1.1 Tarefas

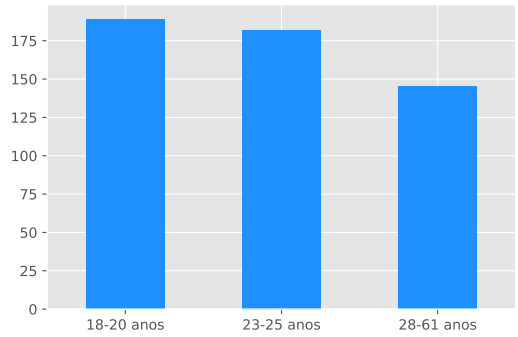
A seguir, é apresentada a definição das tarefas de CA e organização adotada no córpus b5-post, tais como a definição de rótulos para a distribuição de instâncias e a definição do tipo de abordagem de aprendizado supervisionado.

- **Caracterização de faixa etária**: é modelada como um problema de classificação multiclasse (faixas etárias de 18-20, 23-25 e 28-61 anos). Assim como em Rangel et. al. (2015), as classes intermediárias foram desconsideradas como forma de minimizar o erro inerente à classificação de textos publicados ao longo de vários anos por um mesmo autor.
- **Caracterização de gênero e formação em TI**: são modeladas como problemas de classificação binária (masculino/feminino, para gênero; e sim/não, para formação em TI).
- **Caracterização de grau de religiosidade**: é modelada como um problema de classificação multiclasse, em função de sua distribuição no córpus na forma de número inteiro variante de 1 (nenhum pouco religioso) até 5 (muito religioso), foram definidas as classes R1-R2, R3, R4-R5.

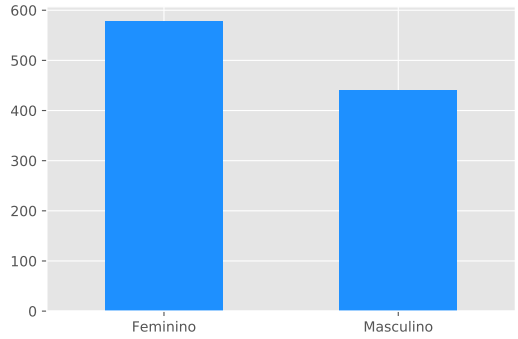
A figura 3 apresenta uma distribuição do número de instâncias por classes de cada uma das tarefas de CA.

Figura 3 – Distribuição de instâncias por tarefas de CA do córpus b5-post

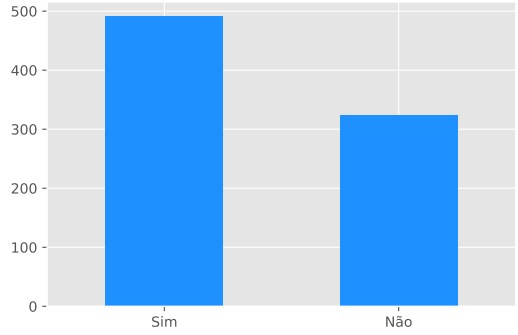
a) Faixa etária



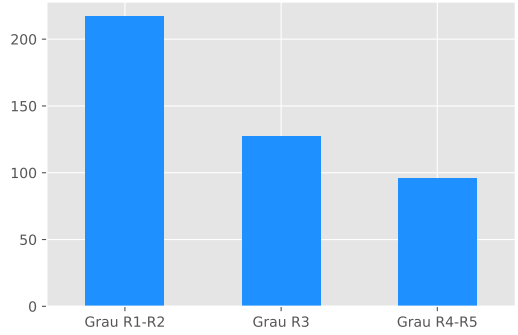
b) Gênero



c) Formação em TI



d) Grau de religiosidade



Fonte: Adaptado de Dias e Paraboni (2018b)

4.1.2 Método

Para o desenvolvimento dos modelos computacionais, foram consideradas combinações de abordagens de aprendizado supervisionado (i.e., classificação de textos) e de representação textual. Os modelos computacionais adotam o classificador categórico de regressão logística, usando as abordagens de acordo com a definição de cada tarefa de CA, e adotam variadas técnicas de representação textual, baseadas em contagem de palavras *Bag-of-Words* (BOW), *n*-gramas de caracteres de tamanho 3 a 5 (Ch-ngram), frequência de palavras (TF-IDF) e vetores distribuídos de palavras (Word2vec) (MIKOLOV et al., 2013a) e sentenças (Doc2Vec) (LE; MIKOLOV, 2014).

Os modelos computacionais BOW, Ch-ngram e TFIDF, foram desenvolvidos usando as implementações da biblioteca em linguagem Python *scikit-learn* (PEDREGOSA et al., 2011). Os vetores distribuídos citados foram treinados a partir de diversas fontes em Português (HARTMANN et al., 2017) usando a arquitetura *Skip-gram*, com janela de

palavras de tamanho 5 e dimensão de tamanho 1000, usando a implementação do algorimo Word2Vec e Doc2vec, da biblioteca em linguagem Python *gensim* (ŘEHŮŘEK; SOJKA, 2010).

O treinamento foi realizado com base no córpus b5-post e utilizado como classificador o algoritmo *liblinear*, da implementação da biblioteca em linguagem Python *scikit-learn* (PEDREGOSA et al., 2011). Foram definidos os hiperparâmetros de regularização L2 e α =1000. A validação foi realizada usando o método de validação cruzada, de 10 partições.

4.1.3 Resultados e discussões

A tabela 1 apresenta os resultados médios de medida F obtidos para cada uma das quatro tarefas organizadas em partes (a-d) e suas respectivas classes. A coluna N é o número de documentos (i.e., autores) considerados em cada tarefa. De modo geral, o modelo TF-IDF apresenta o melhor desempenho exceto para a tarefa de predição de grau de religiosidade, em que o modelo Doc2Vec apresentou resultados melhores. Cabe destacar também os resultados elevados para a tarefa de caracterização de gênero e, em menor escala, para caracterização de formação de TI. No entanto, estes resultados podem estar ligados ao maior volume de dados (N) disponível e não necessariamente ao grau de dificuldade da tarefa. Este experimento inicial será tomado como base para o desenvolvimento de modelos de CA propostos no capítulo seguinte.

4.2 Caracterização de gênero multilíngue (PAN-CLEF 2018)

Em um segundo experimento de CA, consideramos a tarefa de caracterização de gênero multilíngue proposta na tarefa de *Author Profiling* da competição PAN-CLEF 2018 (RANGEL et al., 2018). Para este fim, foram considerados córpus rotulados com informações de gênero, nos idiomas inglês, espanhol e árabe, disponibilizados pela própria competição. O experimento e resultados discutidos a seguir foram apresentados em (DIAS; PARABONI, 2018a).

O córpus utilizado para treinamento foi disponibilizado pela competição PAN-CLEF 2018, contendo publicações (i.e., *tweets*) do Twitter de 10500 autores, segmentados por idioma: 3000 em inglês, 3000 em espanhol e 1500 árabe, e balanceados com 50% de autores

Tabela 1 – Resultados médios de medida F_1 *macro* nas tarefas de CA do córpus b5-post

a) Caracterização de faixa etária

Classes	**N**	**BOW**	**Ch-gram**	**TF-IDF**	**Word2Vec**	**Doc2Vec**
18-20	182	0,51	0,59	0,56	0,57	0,50
23-25	189	0,45	0,48	0,52	0,46	0,47
28-61	145	0,58	0,58	0,65	0,53	0,51

b) Caracterização de gênero

Classes	**N**	**BOW**	**Ch-gram**	**TF-IDF**	**Word2Vec**	**Doc2Vec**
Feminino	578	0,87	0,84	0,90	0,81	0,73
Masculino	440	0,84	0,79	0,86	0,77	0,69

c) Caracterização de formação em TI

Classes	**N**	**BOW**	**Ch-gram**	**TF-IDF**	**Word2Vec**	**Doc2Vec**
Não	491	0,72	0,72	0,75	0,70	0,68
Sim	323	0,60	0,57	0,66	0,57	0,61

d) Caracterização de grau de religiosidade

Classes	**N**	**BOW**	**Ch-gram**	**TF-IDF**	**Word2Vec**	**Doc2Vec**
R1-R2	217	0,57	0,53	0,61	0,60	0,61
R3	96	0,20	0,20	0,28	0,23	0,29
R4-R5	127	0,46	0,40	0,51	0,49	0,55

Fonte: Adaptado de Dias e Paraboni (2018b)

femininos e masculinos. Cada autor com 100 *tweets*. A avaliação foi realizada mediante submissão do modelo ao sistema da competição, que utilizou dados de teste próprios e não acessíveis aos participantes.

4.2.1 Método

Para identificar quais métodos apresentariam melhores resultados para caracterização de gênero multilíngue, foram desenvolvidos cinco modelos computacionais, sendo três de representação textual tradicional: *n*-gramas de palavras, *n*-gramas de caracteres e TF-IDF, e dois de representação distribuída de palavras usando o método de *fasttext* (BOJANOWSKI et al., 2017) baseado em *n*-gramas de caracteres.

Para os modelos de n-gramas foram usadas as sequências de $n \geq 1$ e $n \leq 4$ para palavras e $n \geq 3$ e $n \leq 6$ para caracteres. Para os modelos de representação distribuída foram usados vetores de 300 dimensões pré-treinados com o método de *fasttext* e arquitetura *skip-gram*, a partir de documentos da Wikipédia[1] (BOJANOWSKI et al., 2017) e vetores de 300 dimensões treinados a partir do córpus PAN-CLEF 2018, nos idiomas inglês, espanhol e árabe. O método de *fasttext* considera uma representação distribuída de n-gramas de caracteres, em que cada palavra é representada pela soma de vetores de n-gramas, adotando sequências entre $n \geq 3$ e $n \leq 6$. Este método permite representar uma palavras com base no conjunto de caracteres que a compõem, tornando possível identificar, por exemplo, informações compartilhadas entre palavras, como prefixos, sufixos, radicais e regras morfológicas em geral. Além disso, também possibilita a aproximação de representação de palavras fora do vocabulário (do inglês, *out-of-vocabulary*), por exemplo, palavras raras, variações de tempo verbal ou gírias, podem ser computadas usando a representação do conjunto de caracteres que a compõem.

Os textos do córpus foram organizados por autores e processados utilizando os modelos em questão. Para os modelos de representação distribuída, cada documento (i.e., textos de cada autor) foi representado pela média de vetores de palavras (*word vector averaging*) das publicações (i.e., *tweet*) que compõem cada documento. Cabe destacar que as classes são balanceadas.

A tarefa de caracterização de gênero foi considerada como um problema de classificação binária, e os modelos computacionais adotam o classificador categórico de regressão logística, a partir da implementação de *scikit-learn* (PEDREGOSA et al., 2011), com os hiperparâmetros de regularização L2 e α =1000.

4.2.2 Resultados e discussões

Os modelos foram treinados e testados com o córpus de treinamento, usando validação cruzada k-partes ($k = 10$). Os resultados médios de medida F_1 são apresentados na tabela 2, contendo o idioma, o número de instâncias (N) e os resultados para os modelos: n-gramas de palavras (word n-grams), n-gramas de caracteres (char n-gram), TF-IDF, representação distribuída usando vetores pré-treinados com o método *fasttext* a partir da Wikipédia (*subword*-W), e treinados a partir do córpus de treinamento (*subword*-C).

[1] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Tabela 2 – Resultados médios de medida F_1 para caracterização de gênero do córpus PAN18

Idioma	**N**	**word n-grams**	**char n-grams**	**TF-IDF**	***subword*-W**	***subword*-C**
Inglês	3000	0,80	0,80	0,81	0,67	0,70
Espanhol	3000	0,76	0,76	0,76	0,67	0,67
Árabe	1500	0,75	0,78	0,76	0,68	0,71

Fonte: Adaptado de Dias e Paraboni (2018a)

A avaliação dos modelos computacionais foi realizada pela plataforma TIRA (POTTHAST et al., 2014), e permite que participantes avaliem um único modelo, acessando diretamente a base de teste da competição. Com o objetivo de explorar o método de *fasttext* usando informações de caracteres, o modelo de *subword-W* foi escolhido para ser avaliado no conjunto de teste. A tabela 3 apresenta os resultados de acurácia para a caracterização de gênero, para cada idioma.

Tabela 3 – Resultados de acurácia para caracterização de gênero com base nos dados de teste do córpus PAN18

Idioma	***subword*-W**
Inglês	0,66
Espanhol	0,67
Árabe	0,68

Fonte: Adaptado de Dias e Paraboni (2018a)

4.3 Caracterização de gênero e de usuários robôs no Twitter (PAN-CLEF 2019)

Em um terceiro experimento de CA, considera-se a tarefa de caracterização de gênero e caracterização de usuários robôs no Twitter, proposta na tarefa de *Author Profiling* da competição PAN-CLEF 2019 (RANGEL; ROSSO, 2019). Para este fim, foram considerados córpus textuais rotulados com informações de gênero (i.e., masculino ou feminino) e tipo de perfil (i.e., robô ou humano) nos idiomas inglês e espanhol, disponibilizados pela própria competição. O experimento e resultados discutidos a seguir foram apresentados em (DIAS; PARABONI, 2019).

A tarefa proposta em Rangel e Rosso (2019) consiste em determinar se o autor de um dado texto é um robô ou um humano e, no caso de um humano, determinar o seu gênero. Para esse propósito, dois conjuntos de dados foram disponibilizados, contendo

412.000 instâncias (i.e., documentos) em inglês e 300.000 instâncias em espanhol. Ambos os conjuntos de dados são rotulados com o tipo de perfil (i.e., humano ou robô) e, no caso de perfis de autores humanos, a informação de gênero do autor.

Embora métodos de aprendizado profundo tenham obtido sucesso em diversas áreas de PLN, estes métodos têm sido aplicado em tarefas de CA com nível de sucesso variável. Em particular, observamos que dois dos melhores desempenhos de sistemas na PAN-CLEF 2017 não recorreram a este tipo de método, enquanto que o melhor desempenho na PAN-CLEF-2018 (RANGEL; ROSSO, 2019) usou este tipo de método (TAKAHASHI et al., 2018).

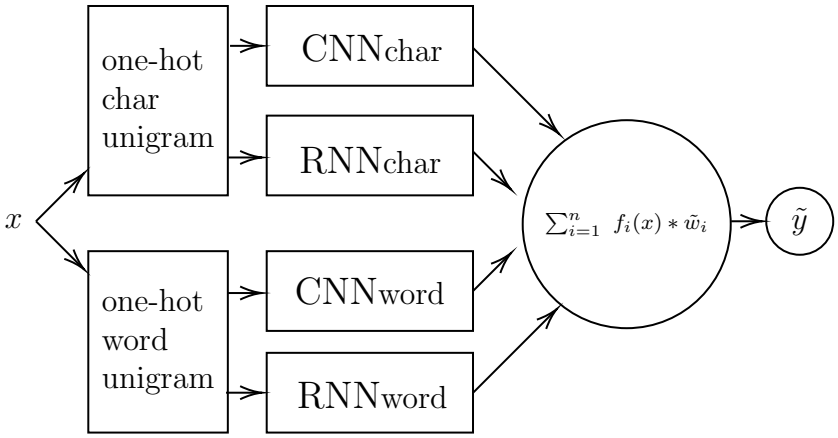
Baseado nestas observações, o presente experimento propõe um método de aprendizado profundo para CA, apresentando uma abordagem de comitê de máquinas com votação ponderada, combinando o aprendizado de redes neurais de convolução (CNN) e recorrência (RNN) baseados em informações de caracteres e palavras. O objetivo é avaliar se a combinação destes modelos tem ganho de desempenho sobre o uso de seus classificadores individualmente, e se o modelo de combinação pode produzir resultados competitivos para a tarefa de caracterização de gênero e caracterização de usuários robôs no Twitter, conforme proposta da PAN-CLEF 2019.

4.3.1 Método

Diferentes estratégias de aprendizado de máquina e representação textual podem promover multiplas contribuições para a tarefa de caracterização de gênero e de usuários robôs. Por isso, o modelo proposto, chamado de *CNN+RNN-char-word*, combina quatro modelos neurais em um comitê de máquinas de votação ponderada, usando a arquitetura de CNN para modelos baseados em informações de nível de caracteres (CNN-char) e de nível de palavras (CNN-word), e arquitetura de RNN para modelos baseados em informações de nível de caracteres (RNN-char) e de nível de palavras (RNN-word). Ambos os modelos de CNNs seguem uma arquitetura multicanal de tamanho de filtro variável, e as RNNs seguem uma arquitetura de LSTM com o mecanismo de atenção (i.e., *self-attention*). A estratégia do comitê de máquinas é ilustrada na figura 4.

Os pesos atribuídos a cada um dos classificadores individuais são calculados a partir de um fator de confiança ótimo para cada modelo, de forma que o erro de saída combinado

Figura 4 – Abordagem de comitê de máquinas CNN+RNN-char-word.



Fonte: Adaptado de Dias e Paraboni (2019)

seja minimizado. A otimização é realizada usando o método *Simplex* (NELDER; MEAD, 1965), usando dados de treinamento de acordo com a equação 14.

$$\tilde{w} = \min \sum_{i} f_i(x) * w_i \tag{14}$$

Nesta equação, para cada modelo i, variamos os pesos de votação $w \in \mathbb{R}$ entre 0 e 1 de forma que minimize a taxa de erro f dos modelos combinados, dado pelo fator de confiança $\tilde{w}$ que representa o peso atribuído para cada modelo nos resultados gerais.

Os textos de entrada são transformados individualmente em vetores de nível de caracteres e de palavras, conforme exigido por cada modelo. Estes vetores são usados para inicialização das camadas de entrada nos modelos CNN e RNN e, consequentemente, otimizados por meio do método de *back-propagation*.

Os modelos de CNN fazem uso de camadas convolucionais de 1 dimensão e operação de agrupamento máximo. Nesta camada são usados os filtros de tamanho 3 e 4, ambos com mapeamentos de tamanho 64 e função de ativação ReLU e regularização L2=$0,003$. Isso é seguido por uma camada totalmente conectada de 1024 neurônios, ativação ReLU, regularização *Dropout*=$0,35$ e camada de saída com ativação *Softmax*. O treinamento foi realizado em lotes de tamanho 32, otimização *RMSProp* e função de custo de entropia cruzada. A validação foi realizada em um conjunto de 20% dos dados de treinamento e método de parada antecipada.

Os modelos da RNN utilizam um mecanismo de atenção do tipo *self-attention*, unidade de memória de tamanho 64 e regularização *Dropout*=$0,12$. Isso é seguido por

Tabela 4 – Resultados de acurácia para a caracterização de usuários robôs usando o conjunto de dados de treinamento da PAN19.

Modelo	Inglês	Espanhol
Baseline (reglog)	0.87	0.83
CNN-char	0.86	0.82
CNN-word	0.67	0.67
RNN-char	0.90	0.86
RNN-word	0.89	0.85
CNN-RNN-char-word	**0**.94	**0**.91

Fonte: Adaptado de Dias e Paraboni (2019)

uma camada oculta de 1024 neurônios usando ativação ReLU e uma camada de saída com ativação *Softmax*. O treinamento é realizado em lotes de tamanho 32, otimização *AdaDelta* e função de custo de entropia cruzada. A validação foi realizada em um conjunto de 20% dos dados de treinamento e método de parada antecipada.

Os resultados dos quatro modelos são combinados usando uma estratégia de votação ponderada, com base em uma estimativa de peso para cada modelo individual de acordo com o desempenho apresentado durante a fase de treinamento.

4.3.2 Resultados e discussões

Esta seção apresenta os resultados da abordagem de comitê de máquinas e sistema de *baseline* aplicado à tarefa de caracterização de gênero e de usuários robôs da PAN-CLEF-2019.

Os resultados de acurácia para a caracterização de usuários robôs são apresentados na tabela 4, e são baseados na avaliação com o conjunto de treinamento da PAN-CLEF-2019. A partir destes resultados, foi observado que a caracterização de usuários robôs foi uma tarefa relativamente simples, apresentando resultados de acurácia acima de 90% para ambos os idiomas. Observa-se também que o comitê de máquinas *CNN-RNN-char-word* geralmente apresenta desempenho superior aos demais modelos, e que o modelo *RNN-char* apresenta o segundo melhor desempenho. Como antecipado na seção anterior, os resultados do modelo de comitê de máquinas são mais influenciados pelo uso de RNNs do que pelo uso de CNNs.

Os resultados de acurácia para a caracterização de gênero são apresentados na tabela 5, e são baseados na avaliação com o conjunto de treinamento da PAN-CLEF 2019.

Tabela 5 – Resultados de acurácia para a caracterização de gênero usando o conjunto de dados de treinamento da PAN19.

Modelo	Inglês	Espanhol
Baseline (reglog)	**0**.74	**0**.72
CNN-char	0.59	0.58
CNN-word	0.55	0.56
RNN-char	0.64	0.63
RNN-word	0.72	0.71
CNN-RNN-char-word	**0**.74	**0**.72

Fonte: Adaptado de Dias e Paraboni (2019)

Tabela 6 – Resultados de acurácia para a caracterização de gênero e de usuários robôs, usando o conjunto de testes da PAN19.

Tarefa	Inglês	Espanhol
Robôs	0.84	0.82
Gênero	0.58	0.65

Fonte: Adaptado de Dias e Paraboni (2019)

Estes resultados sugerem que a caracterização de gênero foi uma tarefa mais desafiadora do que a caracterização de usuários robôs, como evidenciado pelos desempenhos gerais mais baixos se comparados com a tabela 4 anterior. A partir desses resultados, é possível notar que o conjunto *CNN-RNN-char* supera todas as alternativas de modelos de CNN e RNN, em ambos os idiomas, porém os resultados são os mesmos que os obtidos pelo modelo de *baseline LogReg* a um custo computacional muito menor.

Finalmente, a tabela 6 apresenta a avaliação do modelo de comitê de máquinas *CNN-RNN-char-word* para ambas as tarefas baseada no conjunto de teste da PAN-CLEF-2019, cujos resultados foram obtidos na submissão final da tarefa da PAN-CLEF-2019.

Os resultados baseados nos conjuntos de teste são consideravelmente menores do que aqueles observados nos conjuntos de treinamento, e particularmente no caso da tarefa de caracterização de gênero, o que sugere um certo grau de sobreajuste.

Embora o modelo *CNN-RNN-char-word* não tenha obtido uma boa classificação na competição PAN-CLEF 2019, estes resultados nos auxiliaram a definir algumas das estratégias de CA a serem discutidas nos próximos capítulos.

5 Organização dos dados

Este capítulo apresenta a organização dos córpus textuais que foram selecionados para o desenvolvimento desta dissertação e as definições dos problemas abordados. Os córpus são distribuídos em diversos domínios textuais (e.g., redes sociais, blogs, opiniões etc.) e idiomas (e.g., português, inglês e espanhol), e contêm informações anotadas de tarefas de CA (e.g., faixa etária, gênero, grau de escolaridade etc.), representando as classes a serem aprendidas.

A organização destes córpus tem como objetivo a padronização das classes de cada problema de CA. Entretanto, em alguns córpus, é inviável balancear classes sem a perda de informações relavantes. Devido a isso, vamos considerar desbalanceamentos como parte dos problemas de CA abordados nesta dissertação.

Uma visão geral dos córpus é apresentada no quadro 2, contendo o nome dos córpus selecionados, o número de instâncias (i.e., autores), tarefas, idiomas e domínios suportados.

Quadro 2 – Configuração dos córpus

Córpus	Instâncias	Tarefas	Idiomas	Domínios
PAN-CLEF 2013 (EN) (RANGEL et al., 2013)	257.800	G, F	EN	Blogs
PAN-CLEF 2013 (ES) (RANGEL et al., 2013)	109.500	G, F	ES	Blogs
The Blog Authorship (SCHLER et al., 2006)	19.320	G, F	EN	Blogs
b5-post (RAMOS et al., 2018)	1.019	G, F, E, R, T	PT	Facebook
BRMoral (SANTOS; PARABONI, 2019)	433	G, F, E, R, T, P	PT	Opiniões
BlogSet-BR (SANTOS; WOLOSZYN; VIEIRA, 2018)	4.332	G, F, E	PT	Blogs
Nus-SMS (CHEN; KAN, 2013)	44.843	G, F	EN	SMS

Fonte: Rafael Sandroni Dias (2019)

Dois destes córpus são considerados *benchmarks* da área de CA: PAN-CLEF 2013 (RANGEL et al., 2013) e *The Blog Authorship* (SCHLER et al., 2006), enquanto que os córpus BlogSet-BR (SANTOS; WOLOSZYN; VIEIRA, 2018), b5-post (RAMOS et al., 2018), BRMoral (SANTOS; PARABONI, 2019) e Nus-SMS (CHEN; KAN, 2013) estão sendo explorados pela primeira vez para a pesquisa em CA.

As tarefas de CA consideradas são caracterização de gênero (G), faixa etária (F), grau de escolaridade (E), grau de religiosidade (R), formação em Tecnologia da Informação (T) e posição política (P). Os idiomas considerados são o Português (PT), Inglês (EN) e Espanhol (ES).

Os domínios textuais suportados são páginas pessoais da internet (Blogs), redes sociais (Facebook), opinião pessoal (Opinião) e mensagens em serviços de mensagens curtas (SMS).

5.1 Tarefas

São consideradas nove tarefas de CA, distribuídas entre sete córpus textuais, cinco domínios textuais e três idiomas, totalizando 21 problemas de CA. Na tabela 7 é apresentada a relação entre tarefas e córpus, contendo o número de instâncias disponíveis para cada tarefa.

Tabela 7 – Relação entre tarefas e número de instâncias nos córpus

Tarefa	PAN13 EN	PAN13 ES	*The Blog*	b5-post	BRMoral	BlogSet-BR	Nus-SMS
Faixa etária	270.200	109.500	19.243	517	433	2.602	44.843
Gênero	270.200	109.500	19.320	1.019	433	2.602	41.241
Escolaridade					433	1.572	
Religiosidade				441	433		
TI				815	433		
Política					433		

Fonte: Rafael Sandroni Dias (2019)

As tarefas de faixa etária e gênero são suportadas pela totalidade dos córpus selecionados, enquanto que as demais tarefas são suportadas parcialmente, somente por córpus do idioma português (PT).

É possível observar que os córpus PAN-CLEF 2013 e Nus-SMS, no idioma inglês (EN), possuem o maior número de instâncias para a tarefa de gênero, enquanto que os córpus b5-post e BRMoral possuem o menor número de instâncias para a tarefa de faixa etária. Além disso, tarefas menos comuns, tais como escolaridade, religiosidade, TI e política, são suportadas exclusivamente pelos córpus no idioma português.

Na análise da organização destes córpus constatou-se um problema de desbalanceamento de classes. Para minimizar este problema e definir uma padronização, foram realizadas alterações na distribuição das classes dos córpus. Nas subseções seguintes, será discutida detalhadamente cada tarefa, assim como a organização final dos córpus. Nas tabelas seguintes, são apresentadas as informações de classes, total de instâncias (N) e a média de palavras para cada instância (W/N).

5.1.1 Caracterização de faixa etária

A tarefa de caracterização de faixa etária tem como objetivo caracterizar autores com base em grupos de idade. Estes grupos, entretanto, podem variar de acordo com a configuração do córpus. Para esta tarefa, a variável resposta é do tipo categórica, e multiclasse. Entretanto, em alguns córpus a variável resposta pode ser do tipo contínua, representando o valor numérico da idade do autor. A tarefa poderia ter sido modelada como como um problema de regressão.

A organização final dos córpus que suportam a tarefa de faixa etária é apresentada na tabela 8.

Tabela 8 – Organização dos córpus por faixa etária

BlogSet-BR

Classes	N	W/N
a10-25	668	2.409
a26-40	1.020	2.706
a40+	914	4.148
ALL	2.602	3.136

BRMoral

Classes	N	W/N
a0-23	165	311
a24-30	153	303
a31-99	115	289
ALL	433	302

b5-post

Classes	N	W/N
a18-20	182	1.723
a23-25	189	2.092
a28-61	146	2.104
ALL	517	1.966

Blog Authorship

Classes	N	W/N
10s	8.240	5.225
20s	8.086	7.942
30s	2.917	8.559
ALL	19.243	6.872

PAN-CLEF 13 EN

Classes	N	W/N
10s	19.264	731
20s	99.624	550
30s	151.312	725
ALL	270.200	661

PAN-CLEF 13 ES

Classes	N	W/N
10s	4.564	311
20s	56.424	216
30s	48.512	337
ALL	109.500	274

Nus-SMS

Classes	N	W/N
a16-20	22.234	9
a21-30	18.009	10
a35-60	988	15
ALL	41241	10

Fonte: Rafael Felipe Sandroni Dias (2019)

As categorias de faixas etárias são representadas como um conjunto de intervalos abertos entre as idades dos autores, e.g., aM-N sendo M e N os limites inferir e superior. No caso dos córpus PAN-CLEF 2013 e *Blog Authorship*, as classes 10s, 20s e 30s, representam, respectivamente, categorias de faixas etárias com intervalos abertos entre 13-17 anos, 23-27 anos e 33-47 anos.

Os córpus PAN-CLEF 2013, no idioma inglês (EN) e espanhol (ES), *Blog Authorship*, b5-post e BlogSet-BR já possuem organização definida pelos próprios autores e suas definições de faixa etárias foram assim mantidas.

5.1.2 Caracterização de gênero

A tarefa de caracterização de gênero tem como objetivo caracterizar autores com base em seu sexo masculino (M) e feminino (F). A variável resposta é tipo categórica e binária, e a tarefa é considerada como um problema de classificação binária. Nesta tarefa, os córpus não sofreram alterações em suas organizações, conforme apresentado na tabela 9.

5.1.3 Caracterização de grau de escolaridade

A tarefa de caracterização de grau de escolaridade tem como objetivo caracterizar autores com base em níveis de instrução. Estes níveis de instrução variam de acordo com as informações dos córpus. A variável resposta é do tipo categórica e foi padronizada como um problema de classificação multiclasse. Os córpus que suportam esta tarefa passaram por uma reorganização na distribuição de classes. A tabela 10 apresenta a organização final.

As categorias de grau de escolaridade são representadas por valores númericos sucedendo a letra 'S', sendo os valores 0, 1, 2, 3, 4, respectivamente, ensino básico, ensino médio, ensino superior incompleto, ensino superior, e pós-graduação (lato sensu e stricto sensu). Com o objetivo de balancear as classes, agrupamos algumas categorias. Por exemplo, S01 representa o agrupamento das categorias 0 e 1.

Para o córpus BRBlogSet (SANTOS; WOLOSZYN; VIEIRA, 2018), as classes originais "Ensino médio"e "Ensino fundamental" foram agrupadas em S01 (i.e., básico e ensino médio) mantendo-se as demais classes originais.

Tabela 9 – Organização dos córpus por gênero

Blog Authorship

Classes	N	W/N
M	9.660	7.035
F	9.660	6.724
ALL	19320	6.879

BlogSet-BR

Classes	N	W/N
M	1.564	3.531
F	1.038	2.541
ALL	2.602	3.136

BRMoral

Classes	N	W/N
M	285	284
F	148	312
ALL	433	302

b5-post

Classes	N	W/N
M	441	1.873
F	578	2.108
ALL	1019	2.006

PAN-CLEF 2013 EN

Classes	N	W/N
M	135.100	697
F	135.100	625
ALL	270.200	661

PAN-CLEF 13 ES

Classes	N	W/N
M	54.750	274
F	54.750	272
ALL	109.500	273

Nus-SMS

Classes	N	W/N
M	12.718	11
F	28.523	9
ALL	41.241	10

Fonte: Rafael Felipe Sandroni Dias (2019)

Tabela 10 – Organização dos córpus por grau de escolaridade

BlogSet-BR

Classes	N	W/N
S01	303	2.037
S2	341	2.464
S3	465	3.654
S4	463	3.712
ALL	1.572	3.101

BRMoral

Classes	N	W/N
S012	167	318
S3	128	282
S4	138	302
ALL	433	302

Fonte: Rafael Felipe Sandroni Dias (2019)

Para o córpus BRMoral, as classes originais “Fundamental incompleto”, “Fundamental completo”e “Superior incompleto”foram agrupadas em S012, enquanto que as classes S3 (i.e., Superior completo) e S4 (i.e., Pós graduação em andamento ou completa) mantiveram-se na mesma configuração.

5.1.4 Caracterização de grau de religiosidade

A tarefa de caracterização de grau de religiosidade tem como objetivo caracterizar autores com base em graus de religiosidade, de acordo com uma escala pré-definida. O grau de religiosidade é uma variável resposta do tipo categórica, que pode variar de valores de 1 à 5, sendo 1 nada religioso e 5 muito religioso. Por definição, consideramos este um problema de classificação multiclasse.

Pela natureza deste problema, as categorias com valores mais baixos (categoria 1) e valores mais altos (categoria 5) representam dois extremos, i.e., autores que não se consideram religiosos e autores que se consideram muito religiosos Por este motivo, decidimos agrupar as categorias mais extremas e isolar a categoria intermediária (categoria 3), formando três categorias de grau de religiosidade.

Para a organização do córpus b5-post, foram agrupadas as categorias 1 e 2 em R12, representando autores que se consideram nada ou pouco religioso, as categorias 4 e 5 em R45, representando autores que se consideram religiosos ou muito religiosos, e por fim, a categoria 3 em R3, representando autores que se consideram mais ou menos religiosos. Para o córpus BRMoral, a categoria 0 (R0) representa autores nada religiosos, as categorias 3 e 4 (R34) representam autores que se consideram religiosos ou muito religiosos. A tabela 11 apresenta a organização final.

Tabela 11 – Organização dos córpus por grau de religiosidade

BRMoral			b5-post		
Classes	N	W/N	**Classes**	N	W/N
R0	107	317	R12	217	1.827
R12	186	294	R3	96	1.983
R34	140	301	R45	128	2.268
ALL	433	302	ALL	441	1.989

Fonte: Rafael Felipe Sandroni Dias (2019)

5.1.5 Caracterização de formação em TI

A tarefa de caracterização de formação em Tecnologia da Informação (TI) tem como objetivo caracterizar autores que são formados ou trabalham na área de TI. Os córpus

que apresentam esta informação, possuem a variável resposta do tipo categória e binária que por padrão, é considerada como um problema de classificação binária. Portanto, a categoria S representa autores que se formaram ou trabalham na área de TI, e a categoria N representa autores que não se formaram nem trabalham na área de TI. A tabela 12 apresenta a organização final.

Tabela 12 – Organização dos córpus por Formação em TI

BRMoral			b5-post		
Classes	N	W/N	**Classes**	N	W/N
N	140	275	N	491	2.159
S	293	315	S	324	1.796
ALL	433	302	ALL	815	2.014

Fonte: Rafael Felipe Sandroni Dias (2019)

5.1.6 Caracterização de posição política

Finalmente, a tarefa de caracterização de posição política tem como objetivo caracterizar autores com base em posições políticas, de acordo com uma escala pré-definida. Esta tarefa é suportada somente pelo córpus BRMoral que, originalmente, possui variável resposta do tipo categórica e apresenta uma distribuição de categorias de 1 a 5, sendo 1 uma posição política mais à esquerda, 5 uma posiçao política mais à direita. Organizamos a distribuição das categorias para agrupar posições de esquerda (categoria 1) e centro-esquerda (categoria 2) como "P12", posição de direita (categoria 5) e centro-direita (categoria 4) como "P45"e posição de centro (categoria 3) como "P3". Por padrão, este é um problema de classificação multiclasse. A tabela 13 apresenta a organização final.

Tabela 13 – Organização do córpus por posição política

BRMoral		
Classes	N	W/N
P12	157	331
P3	160	287
P45	116	284
ALL	433	302

Fonte: Rafael Felipe Sandroni Dias (2019)

6 Modelos desenvolvidos

Este capítulo apresenta os modelos computacionais desenvolvidos para resolver os problemas de CA apresentados no capítulo anterior. Para este, foram explorados modelos de aprendizado de máquina supervisionado de regressão logística, redes neurais de convolução (CNNs) e redes neurais recorrentes (RNNs), além de técnicas de representação textual de vetores de frequência de palavras, representação distribuída de palavras e vetores de caracteres.

6.1 Visão geral

Foram desenvolvidos sete modelos computacionais para classificação de documentos com objetivo de resolver os problemas de CA listados no capítulo 5. O quadro 3 apresenta uma lista destes modelos e suas principais características.

Quadro 3 – Modelos de CA propostos

Modelo	Classificador	Representação textual
reglog-tfidf	Regressão Logística	TF-IDF
cnn-tfidf	CNN	TF-IDF
cnn-w2v	CNN Multicanal	Word Embeddings
lstm-w2v	LSTM	Word Embeddings
lstm-attention	LSTM + Attention	Word Embeddings
cnn-char	CNN Multicanal	Char Embedding
lstm-char	LSTM + Attention	Char Embedding

Fonte: Rafael Felipe Sandroni Dias (2019)

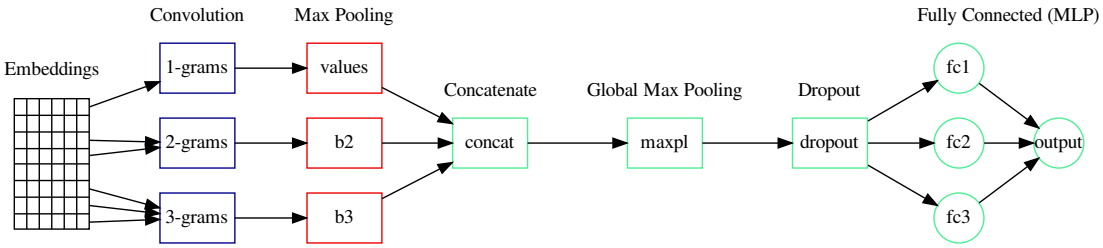
Inicialmente, o modelo *reglog-tfidf* foi definido como o sistema de *baseline*, utilizando um classificador categórico de regressão logística e representação de frequência de palavras TF-IDF. É considerado um modelo padrão da área de PLN e geralmente é uma forma simples e eficiente de resolver problemas de CA (DIAS; PARABONI, 2018b; HSIEH; DIAS; PARABONI, 2018).

O modelo *cnn-tfidf* foi desenvolvido com o objetivo de mesclar representação de frequência de palavras TF-IDF e o aprendizado de uma rede neural de convolução (CNN), para avaliar a capacidade de extração de características da CNN com o uso de informações TF-IDF. A arquitetura definida apresenta uma única camada de convolução e agrupamento

máximo. Os textos foram transformados em vetores de pesos *tf-idf* e então processados pela CNN para extração de características.

Baseado em Kim (2014b), o modelo de *cnn-w2v* apresenta uma arquitetura de CNN multicanal com operações paralelas de convolução com filtros de diferentes tamanhos, e representação distribuída de palavras. É composto por canais de convolução e agrupamento máximo, e uma camada final de neurônios totalmente conectados, como ilustrado na figura 5.

Figura 5 – Ilustração do modelo de CNN-w2v

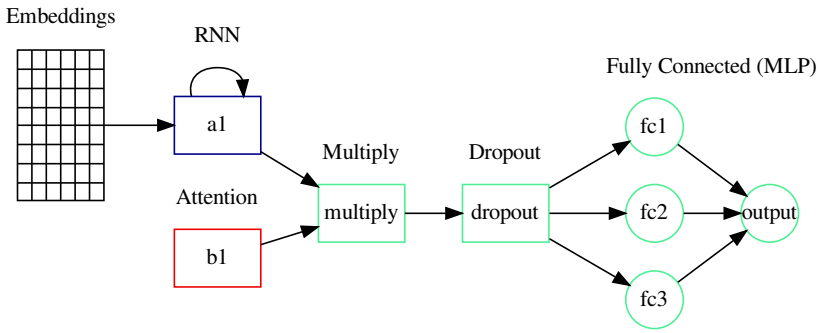


Fonte: Rafael Felipe Sandroni Dias

O modelo *lstm-w2v* utiliza uma arquitetura de LSTM unidirecional e representação distribuída de palavras. Adicionalmente, o modelo de *lstm-attention* apresenta uma arquitetura de LSTM com mecanismo de atenção e representação distribuída de palavras. Este modelo é baseado em Wang et al. (2016) e Bahdanau, Cho e Bengio (2015) e é composto pelos componentes de LSTM unidirecional, mecanismo de atenção e uma camada totalmente conectada, como ilustra a figura 6. A combinação de LSTM e mecanismo de atenção tem o objetivo de identificar quais as características mais importantes aprendidas pela LSTM e então ponderar as características com pesos (*scores*) de atenção. Esta combinação apresentou bons resultados em problemas de CA em Takahashi et al. (2018).

O modelo de *cnn-char* apresenta uma arquitetura de CNN multicanal para classificação de sentenças e utiliza representação de vetores de caracteres. É baseado em Zhang, Zhao e LeCun (2015b) e apresenta resultados do estado da arte em tarefas de PLN. Trabalhar apenas com caracteres permite que combinações anormais de caracteres, como erros de ortografia e emoticons, possam ser aprendidas naturalmente pelo modelo.

Figura 6 – Ilustração do modelo de lstm-attention



Fonte: Rafael Felipe Sandroni Dias

Finalmente, o modelo de *lstm-char* consiste de uma LSTM com mecanismo de atenção, utiliza a arquitetura do modelo *lstm-attention* e representação de vetores de caracteres e é baseado em Zhang, Zhao e LeCun (2015b).

Os modelos desenvolvidos neste capítulo foram treinados e validados usando os conjuntos de treinamento de cada problema de CA. No capítulo seguinte, são apresentadas as avaliações destes modelos com base nos conjuntos de validação.

As implementações foram realizadas usando, especialmente, as bibliotecas *Scikit-learn*, *NLTK*, *gensim* e *tensorflow* da linguagem de programação Python.

6.2 Preparação dos dados

Os córpus textuais passaram por um processo de limpeza, para evitar vocabulário esparso e eliminar possíveis ruídos. Entre as etapas deste processo, estão a remoção de *stopwords*, remoção de etiquetas HTML e transformação em caixa baixa (*lowercase*). Foram utilizados os vocabulários de *stopwords* do repositório NLTK [1], nos idiomas português, inglês e espanhol, de acordo com cada córpus.

Os córpus textuais apresentam, em sua maioria, problemas de desbalanceamentos. Com base em análises sobre experimentos de técnicas de sobreamostragem, a técnica de SMOTE (CHAWLA et al., 2002) foi escolhida para contornar este problema.

[1] http://www.nltk.org

Foram considerados os parâmetros sampling_strategy=*not majority* (i.e., sobreamostragem de todas as classes, exceto a majoritária) e k_neighbors=5 (i.e., número de instâncias vizinhas a serem considerados na construção de exemplos sintéticos para as classes minoritárias).

6.2.1 Representação distribuída de palavras (Word Embeddings)

Foram desenvolvidos modelos de representação distribuída de palavras para cada córpus de CA, os algoritmos considerados foram *Word2Vec* e *FastText*. O treinamento dos modelos foi realizado com o conjunto de treinamento de cada córpus de CA, e selecionado o método de treinamento *skip-gram* e estratégia de amostra negativa (*Negative Sampling*), com os hiperparâmetros de janela de contexto=5 palavras e taxa de aprendizagem (i.e., parâmetro Alpha)=$0,065$, com seleção de palavras com frequência maior que 5.

Foi observado que os vetores treinados com o algoritmo *word2vec* de dimensão dim =100 geralmente apresentam resultados mais estáveis, devido a isso, optamos por considerar apenas modelos treinados nos respectivos córpus e dimensão dim =100.

Foram assim considerados 7 modelos de vetores distribuídos de palavras treinados a partir dos córpus de CA: b5-post, BRMoral, BlogSet-BR, *The Blog Authorship*, PAN 2013 (inglês), PAN 2013 (espanhol) e Nus-SMS.

6.3 Cômputo de hiperparâmetros

Esta seção apresenta a estratégia de treinamento e cômputo de hiperparâmetros dos modelos computacionais. A seguir são apresentadas os hiperparâmetros ótimos encontrados para cada modelo e tarefa de CA, obtidos com base na porção de treinamento de cada córpus.

6.3.1 Modelo de Regressão Logística + TF-IDF (reglog-tfidf)

Os hiperparâmetros do modelo *reglog-tfidf* foram obtidos a partir do método de busca em malha (do inglês, *GridSearch*) com validação cruzada de 10 partições. A tabela

13 apresenta um conjunto pré-definido de valores de hiperparâmetros avaliados e a tabela 5 apresenta os hiperparâmetros ótimos encontrados para cada problema de CA.

Para a representação textual TF-IDF, foram otimizados os hiperparâmetros de *vocab*, limiar máximo de número de palavras no vocabulário, com variação de 1.000 entre 50 (i.e., valor mínimo de palavras) e *max* (i.e., valor máximo dinâmico com base no total de palavras de cada problema de CA), e o hiperparâmetro de *max-idf*, limiar máximo de *idf*, variando entre $0,8$ e $1,0$ a intervalos de $0,1$.

Para o treinamento do modelo, foram otimizados os hiperparâmetros de regularização *reg*, para ajustes no nível de aprendizado do modelo, variando entre L1 e L2, e o hiperparâmetro de Λ, para ajuste dos efeitos de regularização, variando entre $0,0001$ e $10,000$.

O algoritmo de otimização *liblinear* e método de aprendizado OVR (do inglês, *One vs Rest*) foram selecionados para os problemas de classificação multiclasse, enquanto que o algoritmo de otimização $lbfgs$ e método de aprendizado multinomial foram selecionados para os problemas de classificação binária. A tabela 5 apresenta os hiperparâmetros ótimos obtidos para cada problema de CA, *Task* é a tarefa de CA, *Lang* é o idioma e C é o número de classes do problema.

Quadro 4 – Hiperparâmetros considerados para o modelo reglog-tfidf

Hiperparâmetro	Intervalo de valores
Input	$TF-IDF$
max-idf	0,8 - 1,0
max-features	50 - max
Classifier	$logit$
Λ	0,0001 - 10.000
Reg	L1 e L2

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 5 – Hiperparâmetros ótimos para *reglog-tfidf*

PAN 2013

Task	Lang	C	Transformer	max-df	features	Classifier	Λ	Reg
Gender	EN	2	tf-idf	1	661	logit	1000	L2
Gender	ES	2	tf-idf	1	247	logit	1000	L2
Age	EN	3	tf-idf	1	661	logit	1000	L2
Age	ES	3	tf-idf	1	247	logit	1000	L2

The Blog Authorship

Task	Lang	C	Transformer	max-idf	max-F	Classifier	Λ	Reg
Gender	EN	2	tf-idf	1	max	logit	1428,57	L2
Age	EN	3	tf-idf	1	max	logit	1428,57	L2

BRMoral

Task	Lang	C	Transformer	max-idf	max-F	Classifier	Λ	Reg
Gender	PT	2	tf-idf	0,8	max	logit	5714,28	L2
Age	PT	3	tf-idf	0,8	1000	logit	2857,14	L2
Education	PT	3	tf-idf	0,9	1000	logit	10000	L1
IT	PT	2	tf-idf	0,8	max	logit	8571,42	L1
Religion	PT	3	tf-idf	0,8	max	logit	0,0001	L1
Politics	PT	3	tf-idf	0,8	3000	logit	10000	L1

b5-post

Task	Lang	C	Transformer	max-idf	max-F	Classifier	Λ	Reg
Gender	PT	2	tf-idf	0,9	max	logit	1428,57	L2
Age	PT	3	tf-idf	0,8	max	logit	4285,71	L2
IT	PT	2	tf-idf	0,9	max	logit	10000	L1
Religion	PT	3	tf-idf	0,8	max	logit	10000	L2

BlogSet-BR

Task	Lang	C	Transformer	max-idf	max-F	Classifier	Λ	Reg
Gender	PT	2	tf-idf	0,9	max	logit	1428,57	L2
Age	PT	3	tf-idf	1	max	logit	1428,57	L2
Education	PT	4	tf-idf	0,8	max	logit	1428,57	L2

Nus-SMS

Task	Lang	C	Transformer	max-idf	max-F	Classifier	Λ	Reg
Gender	PT	2	tf-idf	0,9	max	logit	1428,57	L2
Age	PT	3	tf-idf	1	max	logit	1428,57	L2

Fonte: Rafael Felipe Sandroni Dias (2019)

6.3.2 Modelo de CNN + TF-IDF (cnn-tfidf)

A tabela 6 apresenta os hiperparâmetros do modelo *cnn-tfidf*, obtidos a partir do domínio do problema de gênero do córpus BR-blogset e então generalizados para os problemas e córpus restantes.

Para a representação textual TF-IDF, o hiperparâmetro de limiar máximo de palavras no vocabulário foi definido pela média de palavras em cada problema de CA *vocab*=*mean*, e o hiperparâmetro de limiar máximo de idf como *max-idf*=$0,9$. Na CNN, os hiperparâmetros da camada de convolução foram defindos como *kernel*=3 e mapeamento de características *feature-maps*=100, com a janela de deslocamento entre palavras *strides*=1 e função de ativação *ReLu*, e hiperparâmetro de agrupamento máximo *pool*=4. O hiperparâmetro de regularização foi definido como *dropout*=0.4, e na camada final *fc*=512 neurônios, com função de ativação *softmax*. O treinamento da CNN foi realizado usando o algoritmo de otimização *AdaDelta* com *learning-rate*=1 e função de custo de Entropia Cruzada, considerando o tamanho do lote *batch-size*=32.

Quadro 6 – Hiperparâmetros assumidos para *cnn-tfidf*

Input	max-idf	vocab	kernel	feature-maps	pool	dropout	fc
TF-IDF	0.9	*mean*	3	100	4	0,4	512

Fonte: Rafael Felipe Sandroni Dias (2019)

6.4 Modelo CNN Multicanal + Word Embeddings (cnn-w2v)

Os hiperparâmetros do modelo *cnn-w2v* foram obtidos a partir do método de busca em malha com validação cruzada de 3 partições. Este número de partições foi definido devido ao elevado tempo de treinamento desse tipo de modelo. A tabela 7 apresenta um conjunto pré-definido de valores de parâmetros avaliados e a tabela 8 apresenta os hiperparâmetros ótimos encontrados para cada problema de CA.

Para a representação textual, foi definido um limiar máximo de palavras no vocabulário *vocab*. Cada palavra foi substituída por um vetor de palavras correspondente e então transformados em uma sequência com limiar máximo *seq* (i.e., por padrão é a média de tokens de cada problema de CA) de vetores de palavras de tamanho *dim*, obtidos a partir dos vetores de palavras do modelo *emb* (i.e., *Word Embeddings*) pré-treinado no

próprio córpus. Quando uma determinada palavra não é encontrada no modelo de vetor de palavras, ela é representada por um vetor de valores zeros de tamanho *dim*. Da mesma forma, quando a instância não possui tamanho *seq*, esta sequência é completada com vetores de valores zeros de tamanho *dim*.

Em seguida, estes vetores são processados pelos canais compostos por camadas de convolução e agrupamento máximo. Neste modelo, os canais são definidos como camadas independentes e os vetores de características resultantes de cada canal são concatenados formando um único vetor, que por sua vez é ajustado pela regularização *dropout*, eliminando conexões menos importantes, e classificado pela camada final de MLP. A classificação é realizada a partir da distribuição probabilística obtida pela função de ativação *Softmax*.

Para o modelo de CNN, foram otimizados os hiperparâmetros de número de canais, indicado por *ch*, e para cada canal, os hiperparâmetros de filtros (*kernels*) k, tamanho de vetores de características (*features maps*) m, agrupamento máximo (*Max Pooling*) de tamanho p, regularização *Dropout* d e número de neurônios na camada final fc. Em especial, na camada de convolução foram definidos os hiperparâmetros de função de ativação *ReLu* e regularização L2 reg=0,3. O treinamento da rede foi realizado com o algoritmo de otimização *AdaDelta* com *learning-rate*=1 e função de custo de entropia cruzada, com *batch-size*=32. A tabela 8 apresenta os hiperparâmetros ótimos encontrados para cada problema de CA para este modelo.

A matriz de entrada recebida pela CNN tem tamanho n x *seq* x *dim*, onde n é o total de instâncias de cada problema de CA, *seq* é o tamanho da sequência de palavras e *dim* é a dimensão do vetor de palavras. Os filtros são de tamanho w x *dim*, w são as janelas deslizantes. Assim, os filtros da camada de convolução realizam a primeira etapa de extração de características no formato de w-gramas.

Quadro 7 – Hiperparâmetros considerados para o modelo cnn-w2v

Hipeparâmetro	Intervalo de valores
vocab	1000 - *max*
seq	50 - *mean*
ch (channel)	1 - 3
k (kernel)	2 - 9
m (feature map)	10 - 100
p (pooling)	1 - 5
d (dropout)	0.1 - 0.5
fc (fully-connected)	256 - 2048

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 8 – Hiperparâmetros ótimos para cnn-w2v

PAN CLEF 2013

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	EN	10000	423	word2vec	100	3	7-8-9	10-10-10	2-2	0.2	512
Age	EN	1000	423	word2vec	100	3	2-3-4	10-10-10	2-2	0.2	512
Gender	ES	10000	409	word2vec	100	3	7-8-9	10-10-10	2-2	0.2	512
Age	ES	1000	409	word2vec	100	2	3-4	40-40	2-2	0.2	512

B5post

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	PT	1000	2069	word2vec	100	2	7-8	10-10	2-2	0.2	512
Age	PT	1000	1957	word2vec	100	3	3-4-5	10-10-10	2-2	0.2	512
IT	PT	1000	1957	word2vec	100	2	7-8	10-10	2-2	0.2	512
Religion	PT	1957	2000	word2vec	100	3	3-4-5	10-10-10	2-2	0.2	512

BRMoral

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	PT	500	302	word2vec	100	2	7-8	50-50	2-2	0.2	512
Age	PT	500	302	word2vec	100	2	7-8	50-50	2-2	0.2	512
Education	PT	500	302	word2vec	100	2	7-8	50-50	2-2	0.2	512
IT	PT	500	302	word2vec	100	2	7-8	50-50	2-2	0.2	512
Religion	PT	1000	302	word2vec	100	2	7-8	50-50	2-2	0.2	512
Politics	PT	6664	302	word2vec	100	2	3	10	2	0.25	512

BRBlogSet

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	PT	1000	3504	word2vec	100	2	3-4	40-40	2-2	0.2	512
Age	PT	1000	3481	word2vec	100	2	7-8	10-10	2-2	0.2	512
Education	PT	15000	3224	word2vec	100	3	3-4	100-100	2-2	0.2	512

Blogs

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	EN	1000	1000	word2vec	100	2	3-4	10-10	2-2	0.2	512
Age	ES	15000	3897	word2vec	100	2	3-4	50-50	2-2	0.2	512

SMS

task	lang	vocab	seq	emb	dim	ch	k	m	p	d	fc
Gender	EN	4309	50	word2vec	100	2	3-4-5	10-10	2-2	0.2	512
Age	EN	4171	50	word2vec	100	2	2-3	10-10	2-2	0.2	512

Fonte: Rafael Felipe Sandroni Dias (2019)

6.5 Modelo LSTM + Word Embeddings (lstm-w2v)

Os hiperparâmetros do modelo *lstm-w2v* foram obtidos a partir do problema de caracterização de gênero do córpus BRBlogSet, com validação cruzada de 3 partições. Esta estratégia foi definida devido ao elevado tempo de treinamento desse tipo de modelo. A tabela 10 apresenta os hiperparâmetros ótimos.

Para a representação textual, foi definido um limiar máximo de palavras no vocabulário *vocab*=*max* (i.e., total de palavras de cada problema de CA). Cada palavra foi substituída por um vetor de palavras correspondente, e então transformados em uma sequência com limiar máximo *seq*=*mean* (i.e., média de palavras de cada problema de CA) de vetores de palavras de tamanho *dim*, obtidos a partir do modelo *emb* (i.e., *Word Embeddings*). Quando uma determinada palavra não é encontrada no modelo de vetor distribuído de palavras, ela é representada por um vetor de valores zeros de tamanho *dim*, assim como quando a instância não possui tamanho *seq*, esta sequência é completada com vetores de valores zeros de tamanho *dim*.

Para o modelo de LSTM, foi definido o hiperparâmetro de unidades de memória *units*=128, regularização *dropout* d=0,2 e número de neurônios na camada final fc=512. O treinamento da rede é realizado com o algoritmo de otimização *RMSProp* com *learning-rate*=1 e função de custo de entropia cruzada, com *batch-size*=32. A tabela 10 apresenta os hiperparâmetros ótimos generalizados para os problema de CA.

Quadro 9 – Hiperparâmetros considerados para o modelo lstm-w2v

Hipeparâmetro	Intervalo de valores
vocab	1000 - *max*
seq	*mean*
units	64 - 256
dropout	0.1 - 0.5
fc	512 - 2048

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 10 – Hiperparâmetros ótimos para *lstm-w2v*

vocab	seq	embed	dim	units	attention-n	dropout	fc
max	*mean*	word2vec	100	128	*seq*	0.2	512

Fonte: Rafael Felipe Sandroni Dias (2019)

6.6 LSTM com Mecanismo de Atenção + Word Embedding (lstm-attention)

Os hiperparâmetros do modelo *lstm-attention* foram obtidos a partir do problema de caracterização de gênero do córpus BRBlogSet, com validação cruzada de 3 partições. Novamente, esta estratégia foi definida devido ao elevado tempo de treinamento desse tipo de modelo. A tabela 12 apresenta os hiperparâmetros ótimos deste modelo.

Para a representação textual, o tratamento foi o mesmo aplicado nos modelos *lstm-w2v* e *cnn-w2v* anteriores.

Para o modelo de LSTM, foi definido o hiperparâmetro de unidades de memória *units*=128, os valores de atenção são obtidos a partir dos estados ocultos da LSTM. Para essa etapa, foi utilizado uma MLP com número de neurônios *attention-n*=*seq*, e em sequida, os hiperparâmetros de regularização *dropout* d=0,2 e número de neurônios na camada final fc=512. O treinamento da rede é realizado com o algoritmo de otimização *RMSProp* com *learning-rate*=1 e função de custo de entropia cruzada, com *batch-size*=32. A tabela 16 apresenta os hiperparâmetros generalizados para os problema de CA.

Quadro 11 – Hiperparâmetros considerados para o modelo lstm-attention

Hipeparâmetro	Intervalo de valores
vocab	1000 - *max*
seq	50 - *mean*
units	64 - 256
dropout	0,1 - 0,5
fc	512 - 2048

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 12 – Hiperparâmetros ótimos para *lstm-attention*

vocab	seq	embed	dim	units	attention-n	dropout	fc
max	*mean*	word2vec	100	128	*seq*	0.2	512

Fonte: Rafael Felipe Sandroni Dias (2019)

6.7 CNN Multicanal + Char (cnn-char)

Os hiperparâmetros do modelo *cnn-char* também foram obtidos a partir do problema de caracterização de gênero do córpus BRBlogSet, com validação cruzada de 3 partições,

devido ao elevado tempo de treinamento desse tipo de modelo. A tabela 14 apresenta os hiperparâmetros ótimos obtidos.

A representação textual é realizada dividindo o texto por sentenças, cada caracter é representado por um vetor de tamanho *dim*=72, com um limiar máximo de caracteres por sentença *seq*=*char mean* (i.e., média de caracteres por sentença). O vocabulário de caracteres é ilustrado abaixo.

```
abcdefghijklmnopqrstuvwxyz0123456789-,;.!?:'\"/\\|_@#$%^&*~`+-=<>()[]{}
```

O modelo de CNN utiliza dois canais de convolução *ch*=2 com filtros de tamanho *kernel*=3-4, ambos canais com mapeamento de tamanho *map*=64, usando a função de ativação *relu* e regularização L2 de reg=$0,003$ e agrupamento máximo de tamanho *pool*=2. Seguido de uma camada totalmente conectada de fc=1024 neurônios usando função de ativação *relu* e regularzação *Dropout* (SRIVASTAVA et al., 2014) de d=$0,3$, e ao final uma camada de saída usando função *softmax*. O treinamento é realizado em mini-lotes de tamanho *batch*=32 usando o algoritmo de otimização RMSProp e função de custo de entropia cruzada. A validação do treinamento é realizada com 20% dos dados do conjunto de treinamento e executada até o modelo convergir, usando a técnica de *EarlyStopping*.

Quadro 13 – Hiperparâmetros considerados para o modelo cnn-char

Hiperparâmetro	Intervalo de valores
vocab	27
seq	*char mean*
ch	2 - 3
k	2 - 9
map	10 - 100
d	0,1 - 0,5
fc	512 - 2048

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 14 – Hiperparâmetros ótimos para *cnn-char*

max features	seq	embed.	dim	ch	k	map	d	fc
max	*mean*	chars	27	2	3-4	64	0.2	512

Fonte: Rafael Felipe Sandroni Dias (2019)

6.8 LSTM com Mecanismo de Atenção + Char (lstm-char)

Os hiperparâmetros do modelo *lstm-char* também foram obtidos a partir do problema de caracterização de gênero do córpus BRBlogSet, com validação cruzada de 3 partições, devido ao elevado tempo de treinamento desse tipo de modelo. A tabela 15 apresenta os hiperparâmetros ótimos obtidos.

A representação textual é realizada dividindo o texto por sentenças, cada caracter é representado por um vetor de tamanho *dim*=27, com um limitar máximo de caracteres por sentença *seq*=*char mean* (i.e., média de caracteres por sentença). O vocabulário considerado é ilustrado abaixo.

```
abcdefghijklmnopqrstuvwxyz0123456789-,;.!?:'\"/\\|_@#$%^&*~`+-=<>()[]{}
```

O modelo de *lstm-char* utiliza unidades de memória de tamanho *units*=128 e regularização *Dropout* de d=0, 2, seguido de uma camada oculta de fc=512 neurônios com função de ativação *ReLu* e uma camada de saída usando *softmax*. O treinamento é realizado usando o algoritmo de otimização AdaDelta e função de custo de entropia cruzada. A validação do treinamento é realizado com 20% dos dados do conjunto de treinamento e executada até o modelo convergir, usando a técnica de *EarlyStopping*.

Quadro 15 – Hiperparâmetros considerados para o modelo lstm-char

Hipeparâmetro	Intervalo de valores
vocab	27
seq	*char mean*
units	64 - 128
d	0.1 - 0.5
fc	512 - 2048

Fonte: Rafael Felipe Sandroni Dias (2019)

Quadro 16 – Hiperparâmetros ótimos para *lstm-char*

vocab	seq	embed	dim	units	attention-n	dropout	fc
max	*char mean*	chars	27	64	*seq*	0.12	512

Fonte: Rafael Felipe Sandroni Dias (2019)

7 Avaliação

Neste capítulo, são apresentados os resultados gerais dos modelos computacionais propostos, baseados nos conjuntos de teste dos respectivos córpus de CA. Os modelos computacionais desenvolvidos foram aplicados para cada um dos 21 problemas de CA, totalizando 147 conjuntos de resultados. Nas seções seguintes, os resultados são organizados por tarefas, e para cada tarefa são apresentados os resultados de medida F (F_1 *score*), e os melhores desempenhos são destacados.

Nesta avaliação, as métricas foram computadas usando a média-macro, mantendo-se pesos iguais para todas as classes e as classes foram balanceadas com o método SMOTE de sobreamostragem (CHAWLA et al., 2002). A tarefa de caracterização de gênero dos córpus *The Blog Authorship* e PAN-CLEF 13 não receberam o tratamento de sobreamostragem.

É importante destacar que o modelo *reglog-tfidf*, descrito no capítulo 6, será demoninado neste capítulo como simplesmente *baseline* para facilitar a análise comparativa.

7.1 Caracterização de gênero

A tabela 14 apresenta os resultados obtidos para a tarefa de caracterização de gênero.

Tabela 14 – Resultados gerais de medida F_1 para caracterização de gênero

Córpus	*baseline*	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
Nus-SMS	0,74	0,55	0,33	0,74	**0,75**	0,72	0,71
PAN13 EN	0,56	0,57	0,41	0,51	**0,64**	0,46	0,35
PAN13 ES	0,51	0,44	0,33	0,50	**0,60**	0,42	0,54
b5-post	**0,86**	0,82	0,78	0,65	0,81	0,48	0,44
BlogSet-BR	0,76	0,78	0,62	0,72	0,73	**0,80**	**0,80**
BRMoral	0,58	0,53	**0,67**	0,41	0,66	0,59	0,58
The Blog	**0,78**	0,65	0,75	0,72	**0,78**	0,53	0,55

Fonte: Rafael Felipe Sandroni Dias (2019)

Com base nos resultados gerais obtidos, é possível observar que modelos de *lstm-attention* apresentam maior quantidade de resultados positivos. Apenas no córpus *b5-post* o modelo de *baseline* superou os demais modelos, enquanto que no córpus *The Blog Authorship*, houve um empate entre o modelo de *LSTM-attention* e *baseline*.

7.2 Caracterização de faixa etária

A tabela 15 apresenta os resultados obtidos para a tarefa de caracterização de faixa etária.

Tabela 15 – Resultados gerais de medida F_1 para caracterização de faixa etária

Córpus	*baseline*	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
Nus-SMS	0,70	0,46	0,56	0,66	0,69	**0,71**	0,43
PAN13 EN	0,50	0,57	0,49	0,51	**0,65**	0,55	0,49
PAN13 ES	0,38	0,35	0,30	0,44	0,60	0,28	**0,61**
b5-post	0,58	**0,62**	0,32	0,17	0,27	0,31	0,29
BlogSet-BR	0,28	**0,45**	0,39	0,37	0,37	0,36	0,34
BRMoral	**0,41**	0,37	0,17	0,30	0,38	0,36	0,37
The Blog	0,75	0,73	**0,76**	0,72	0,69	0,35	0,35

Fonte: Rafael Felipe Sandroni Dias (2019)

Os modelos computacionais desenvolvidos para a tarefa de caracterização de faixa etária apresentam resultados de níveis variados e distribuídos entre todas alternativas. Cada modelo obteve resultados superiores em pelo menos um córpus, com exceção do modelo *lstm-w2v*. O modelo *cnn-tfidf* apresentou desempenhos superiores em dois córpus, *b5-post* e *BlogSet-BR*. Nenhum dos modelos propostos apresentou resultado superior ao *baseline* no córpus BRMoral.

7.3 Caracterização de grau de escolaridade

A tabela 16 apresenta os resultados obtidos para a tarefa de caracterização de grau de escolaridade.

Tabela 16 – Resultados gerais de medida F_1 para caracterização de escolaridade

Córpus	baseline	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
BlogSet-BR	0,35	**0,37**	0,31	0,25	0,10	0,21	0,21
BRMoral	**0,35**	0,34	0,28	0,32	0,31	0,29	0,29

Fonte: Rafael Felipe Sandroni Dias (2019)

Os resultados gerais da tarefa de caracterização de grau de escolaridade sugerem que esta é uma tarefa difícil. Apenas o modelo *cnn-tfidf* apresentou desempenho superior ao *baseline* no córpus BlogSet-BR. Nenhum dos modelos propostos apresentou resultado superior ao *baseline* no córpus BRMoral.

7.4 Caracterização de grau de religiosidade

A tabela 17 apresenta os resultados obtidos para a tarefa de caracterização grau de religiosidade.

Tabela 17 – Resultados gerais de medida F_1 para caracterização de religiosidade

Córpus	baseline	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
b5-post	0,36	0,42	**0,56**	0,17	0,33	0,44	0,33
BRMoral	**0,42**	0,35	0,32	0,38	0,27	0,27	0,30

Fonte: Rafael Felipe Sandroni Dias (2019)

Com base nos resultados gerais, é possível observar que o modelo de *cnn-w2v* apresenta resultados superiores no córpus *b5-post*, com 20 pontos acima do modelo de *baseline*. Nenhum dos modelos propostos apresentou resultado superior ao *baseline* no córpus BRMoral.

7.5 Caracterização de formação em TI

A tabela 18 apresenta os resultados obtidos para a tarefa de caracterização de formação em TI.

Tabela 18 – Resultados gerais de medida F_1 para caracterização de formação em TI

Córpus	baseline	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
b5-post	**0,71**	0,63	0,61	0,60	0,55	0,41	0,43
BRMoral	0,60	0,53	0,72	0,71	0,43	**0,74**	0,69

Fonte: Rafael Felipe Sandroni Dias (2019)

É possível observar que os modelos computacionais desenvolvidos para a tarefa de caracterização de formação em TI apresenta resultados gerais variados. No córpus *b5-post*, por exemplo, o *baseline* superou os modelos propostos. Os modelos baseados em informações de *char* apresentaram os piores desempenhos, enquanto que os mesmos modelos apresentaram resultados superiores ao *baseline* no córpus BRMoral.

7.6 Caracterização de posição política

A tabela 19 apresenta os resultados obtidos para a tarefa de caracterização de posição política.

Tabela 19 – Resultados gerais de medida F_1 para caracterização de posição política

Córpus	baseline	CNN-tfidf	CNN-w2v	LSTM-w2v	LSTM-attention	CNN-char	LSTM-char
BRMoral	0,49	**0,53**	0,39	0,45	0,17	0,23	0,24

Fonte: Rafael Felipe Sandroni Dias (2019)

Os resultados gerais da tarefa de caracterização de posição política destacam um nível de variação de desempenho entre os modelos propostos. O modelo *cnn-tfidf* apresentou o melhor resultado, e o sistema de *baseline* apresenta o segundo melhor resultado.

7.7 Considerações

Com base nos desempenhos apresentados, é possível observar que os resultados dos modelos propostos são mistos, e muitas vezes próximos do *baseline*. A tabela 17 ilustra um resumo geral dos modelos com os melhores desempenhos, organizados por córpus e por tarefa.

Quadro 17 – Resumo geral dos melhores desempenhos, por córpus e tarefa

	Tarefas					
Córpus	Gênero (G)	Faixa etária (F)	Escolaridade (E)	Religiosidade (R)	TI (T)	Política (P)
PAN13 EN	LSTM-attention	LSTM-attention				
PAN13 ES	LSTM-attention	LSTM-char				
Nus-SMS	LSTM-attention	CNN-char				
The Blog	**baseline**/LSTM-attention	CNN-w2v				
BlogSet-BR	CNN-char/LSTM-char	CNN-tfidf	CNN-tfidf			
b5-post	**baseline**	CNN-tfidf		CNN-w2v	**baseline**	
BRMoral	CNN-w2v	**baseline**	**baseline**	**baseline**	CNN-char	CNN-tfidf

Fonte: Rafael Felipe Sandroni Dias (2019)

A tarefa de caracterização de gênero (G) apresenta resultados variados entre os modelos baseados em CNN e LSTM, e o modelo de *baseline*. Obverva-se que os modelos de RNAs foram mais eficientes na maioria dos córpus, enquanto que o modelo de *baseline* superou os modelos de RNAs apenas no córpus b5-post, e houve um empate entre baseline e *LSTM-attention* no córpus *The Blog Authorship*. Dentre os resultados mais significativos, o modelo de *LSTM-attention* obteve 13 pontos acima do *baseline* na tarefa de caracterização de gênero (G) do córpus PAN13 ES; o modelo de *CNN-w2v* obteve 9 pontos acima do *baseline* na tarefa do córpus BRMoral; e os modelos de *CNN-char* e *LSTM-char*, alcançaram 4 acima do modelo de *baseline* na tarefa do córpus BlogSet-BR .

Na tarefa de caracterização de faixa etária (F), o uso de RNAs apresenta desempenho superior ao *baseline*. De forma mais específica, os modelos de CNN demonstraram-se mais eficientes em 57% das vezes (i.e, 4 em 7). Dentre os resultados mais significativos, o modelo

LSTM-char obteve 23 acima do baseline na tarefa (F) do córpus PAN13 ES; o modelo de *CNN-tfidf* alcançou 17 acima do baseline na tarefa (F) do córpus BlogSetBR; e por fim, o modelo *LSTM-attention* obteve 15 pontos acima do modelo de baseline na tarefa (F) do córpus PAN13 EN.

As tarefas de caracterização de grau de escolaridade (E), grau de religiosidade (R), formação em TI (T) apresentam resultados mistos, com os modelos de CNNs e o *baseline*. O modelo de *CNN-tfidf* teve o melhor desempenho na tarefa de caracterização de posição política (P). Dentre os resultados mais significativos, o modelo de *CNN-w2v* obteve 20 pontos acima do baseline na tarefa de caracterização de religiosidade (R); e o modelo *CNN-char* obteve 14 pontos acima do baseline na tarefa de caracterização de formação em TI (T) do córpus BRMoral.

Dentre os modelos de RNAs, os modelos baseados em CNNs apresentaram desempenho superior em 10 das 21 tarefas, enquanto que as LSTMs obtiveram desempenho superior em 7 das 21 tarefas. O modelo de *LSTM-attention* obteve sucesso em maior quantidade de tarefas, totalizando 5 tarefas. Estes desempenhos são observados em córpus de tamanho médio e grande, como PAN13-EN, PAN13-ES, Nus-SMS e *The Blog Authorship*. O mecanismo de atenção demonstrou-se capaz de melhorar os resultados da LSTM com a extração de características de contexto, como a relação e dependência semântica entre palavras. Adicionalmente, alguns modelos de CNN, principalmente o modelo de *CNN-tfidf* e *CNN-char* também obtiveram resultados positivos em córpus pequenos, como BRMoral, b5-post e BlogSet-BR. Estes modelos demonstraram-se capazes de realizar a extração de características semânticas, como identificação padrões de estrutura de palavras, de forma mais eficiente em córpus menores.

Para ilustrar esta relação entre tamanho do córpus e desempenho, o quadro 18 apresenta um resumo geral de quais tarefas foram superadas por RNAs e *baseline*, ordenando pelo número de instâncias (N) de cada córpus. Os dados demonstram que o modelo de *baseline* não é competitivo em córpus grandes.

7.7.1 Comparativo com os resultados oficiais da PAN-CLEF 2013

Finalmente, os resultados gerais dos modelos propostos de *lstm-attention* e *lstm-char* nos córpus PAN13 EN e PAN13 ES foram comparados com os resultados oficiais do vencedor da competição PAN-CLEF 2013 (RANGEL et al., 2013). O modelo em Meina et al. (2013b) foi o vencedor da competição obtendo o melhor resultado combinado na competição (i.e., média dos resultados das tarefas de caracterização de gênero e faixa etária nos idiomas inglês e espanhol). A tabela 20 resume os resultados de acurácia obtidos no conjunto de testes dos córpus PAN13 EN

Quadro 18 – Resumo geral dos melhores resultados por córpus, ordenado pelo número de instâncias (N)

Córpus	**N** ↑	Avg words	RNAs	Baseline
PAN13 EN	270.200	661	G, F	
PAN13 ES	109.500	274	G, F	
Nus-SMS	41.241	10	G, F	
The Blog	19.243	6.872	F, G	G
BlogSet-BR	2.602	3.136	G, F, E	
b5-post	1019	1.966	F, R	G, T
BRMoral	433	302	G, T, P	F, E, R

Fonte: Rafael Felipe Sandroni Dias (2019)

e PAN13 ES. O modelo de *LSTM-attention* obteve 4 e 1 pontos acima do modelo de Meina et al. (2013b), respectivamente, nas tarefas de caracterização de gênero e de faixa etária do córpus PAN13 EN, e 10 pontos acima na tarefa de caracterização de faixa etária do córpus PAN13 ES, enquanto que o modelo de *LSTM-char* obteve 5 pontos acima na tarefa de caracterização de gênero do córpus PAN13 ES.

Tabela 20 – Resultados de acurácia obtidos no conjunto de testes do córpus PAN13 EN e PAN13 ES

Córpus	Tarefa	LSTM-attention	LSTM-char	(MEINA et al., 2013b)
PAN13 EN	Gênero	0,63	0,35	0,59
	Faixa Etária	0,65	0,48	0,64
PAN13 ES	Gênero	0,58	0,54	0,53
	Faixa Etária	0,58	0,59	0,49

Fonte: Rafael Felipe Sandroni Dias (2019)

8 Conclusão

Este trabalho apresentou uma pesquisa em nível de mestrado dedicado às tarefas de caracterização autoral (CA) a partir de textos utilizando redes neurais artificiais (RNAs). A pesquisa consistiu em organizar os córpus textuais, definir as tarefas de CA e implementar uma série de modelos computacionais utilizando abordagens de classificação de documentos. Este trabalho apresentou as seguintes contribuições.

- Organização dos córpus e definição de tarefas de CA
- Modelos computacionais propostos
- Resultados de referência para futuros estudos da área
- Resultados superiores ao estado-da-arte da competição PAN 2013

Além do estudo de tarefas de caracterização de gênero e faixa etária, muito comuns em pesquisas de CA, este trabalho teve como contribuição a exploração de tarefas possivelmente inéditas em CA, tais como caracterização de grau de religiosidade, formação em TI e posição política no idioma português.

A pesquisa explorou o uso de modelos computacionais baseados em aprendizado profundo e foram adotadas diversas abordagens e arquiteturas de RNAs que apresentaram resultados promissores na área de PLN (KIM, 2014b; WANG et al., 2016; ZHANG; ZHAO; LECUN, 2015b). Os modelos baseados em *Convolutional Neural Network* (CNN) apresentaram a maior combinação de resultados positivos em diversas tarefas de CA. Os modelos baseados em *Long Short Term Memory* (LSTM) apresentaram desempenho superior em córpus de maior volume de dados, especialmente nos córpus PAN13, Nus-SMS e *The Blog Authorship*. Os modelos de *LSTM-attention* e *LSTM-char* superaram os resultados de estado-da-arte da competição PAN-CLEF 2013 nas tarefas de caracterização de gênero e caracterização de faixa etária, nos idiomas inglês e espanhol.

Os modelos propostos e seus resultados serão objetivo de futuras publicações e servem como base para futuros estudos na área de CA, tais modelos estão disponíveis no repositório do autor no *github.com* [1]. Os córpus foram organizados e separados entre conjuntos de treinamento e teste, e podem ser obtidas mediante solicitação ao autor.

[1] https://github.com/rafaelsandroni/author-profiling-models

8.1 Publicações derivadas deste trabalho

1. **Caracterização autoral de usuários do Facebook brasileiro (DIAS; PARABONI, 2018b; HSIEH; DIAS; PARABONI, 2018)**: Este artigo, descrito em detalhes no capítulo 4 seção 4.1, teve como objetivo a organização dos córpus textuais, assim como o desenvolvimento inicial de modelos computacionais para as tarefas de CA. Destes modelos, foram exploradas técnicas de representação textual de vetores de contagem de palavras (*Bag-of-Words*) à representação distribuída de palavras (*Word Embeddings*).
2. **Author Profiling using Word Embeddings with Subword Information (PAN-CLEF 2018) (DIAS; PARABONI, 2018a)**: Este artigo, descrito em detalhes no capítulo 4 seção 4.2, teve como objetivo explorar abordagens de representação distribuída de palavras baseadas em informações de *subwords* (*FastText*). Essa publicação é resultado desses experimentos e da participação da competição internacional de CA PAN-CLEF 2018 (RANGEL et al., 2018).
3. **Combined CNN+RNN Bot and Gender Profiling (PAN-CLEF 2019) (DIAS; PARABONI, 2019)**: Neste artigo, descrito em detalhes no capítulo 4 seção 4.3, é apresentado experimentos envolvendo o uso de CNNs e RNNs para tarefas de CA. Foram desenvolvidos modelos baseados em caracteres para a competição PAN-CLEF 2019, e em certo momento deste projeto, foram realizados experimentos combinando os modelos desenvolvidos em um comitê de máquinas com votação ponderada. Essa publicação é resultado desses experimentos e da participação da competição internacional de CA PAN-CLEF 2019 (RANGEL; ROSSO, 2019).

8.2 Outras colaborações

1. **Building a Corpus for Personality-dependent Natural Language Understanding and Generation (RAMOS et al., 2018)** Este artigo descreve a estrutura do córpus b5-post utilizado neste trabalho, e cuja construção foi realizada com a participação do autor.

Referências[2]

ÁLVAREZ-CARMONA, M. et al. INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In: *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. [S.l.]: CEUR-WS.org, 2015. ISSN 1613-0073. Citado na página 31.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 26 e 75.

BARTLE, A.; ZHENG, J. *Gender classification with deep learning*. [S.l.]: Stanford Technical Report, 2015. Citado 3 vezes nas páginas 45, 53 e 55.

BASILE, A. et al. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*. Dublin, Ireland: CEUR-WS.org, 2017. ISSN 1613-0073. Citado 4 vezes nas páginas 18, 23, 32 e 55.

BENGIO, Y. et al. A neural probabilistic language model. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 1137–1155, mar. 2003. ISSN 1532-4435. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=944919.944966⟩. Citado na página 22.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, IEEE, v. 5, n. 2, p. 157–166, 3 1994. Citado na página 25.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Association for Computational Linguistics, v. 5, p. 135–146, 2017. Citado 2 vezes nas páginas 60 e 61.

BURGER, J. D. et al. Discriminating gender on twitter. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 1301–1309. ISBN 978-1-937284-11-4. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2145432.2145568⟩. Citado 2 vezes nas páginas 16 e 44.

CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, v. 2380 de *CEUR Workshop Proceedings*, (CEUR Workshop Proceedings, v. 2380). CEUR-WS.org, 2019. Disponível em: ⟨http://ceur-ws.org/Vol-2380⟩. Nenhuma citação no texto.

CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, v. 2125 de *CEUR Workshop Proceedings*, (CEUR Workshop Proceedings, v. 2125). CEUR-WS.org, 2018. Disponível em: ⟨http://ceur-ws.org/Vol-2125⟩. Nenhuma citação no texto.

CARMONA, M. A. A. et al. Inaoe's participation at PAN'15: Author profiling task. In: *CLEF (Working Notes)*. Toulouse, France: CEUR-WS.org, 2015. v. 1391. Citado na página 55.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 16, n. 1, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=1622407.1622416⟩. Citado 2 vezes nas páginas 76 e 88.

[2] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CHEN, T.; KAN, M.-Y. Creating a live, public short message service corpus: The NUS SMS corpus. *Language Resources and Evaluation*, v. 47, n. 2, p. 299–335, 2013. Citado na página 67.

CHO, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *EMNLP*. ACL, 2014. p. 1724–1734. ISBN 978-1-937284-96-1. Disponível em: ⟨http://dblp.uni-trier.de/db/conf/emnlp/emnlp2014.html#ChoMGBBSB14⟩. Citado na página 27.

CHUNG, J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: *NIPS 2014 Workshop on Deep Learning, December 2014*. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 14 e 27.

CONNEAU, A. et al. Very deep convolutional networks for natural language processing. *KI - Künstliche Intelligenz.*, v. 26, 06 2016. Citado na página 14.

DIAS, R. F. S.; PARABONI, I. Author profiling using word embeddings with subword information: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.* CEUR-WS.org, 2018. (CEUR Workshop Proceedings, v. 2125). Disponível em: ⟨http://ceur-ws.org/Vol-2125/paper_97.pdf⟩. Citado 4 vezes nas páginas 56, 59, 62 e 95.

DIAS, R. F. S.; PARABONI, I. Caracterização autoral de usuários do facebook: gênero, idade, religiosidade e área de formação. In: *I Congresso Internacional em Humanidades Digitais (HDRio-2018)*. Rio de Janeiro: [s.n.], 2018. Citado 5 vezes nas páginas 56, 58, 60, 74 e 95.

DIAS, R. F. S.; PARABONI, I. Combined CNN+RNN bot and gender profiling. In: CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019.* CEUR-WS.org, 2019. (CEUR Workshop Proceedings, v. 2380). Disponível em: ⟨http://ceur-ws.org/Vol-2380/paper_61.pdf⟩. Citado 6 vezes nas páginas 56, 62, 64, 65, 66 e 95.

FATIMA, M. et al. Multilingual author profiling on facebook. *Information Processing and Management*, Elsevier, v. 53, n. 4, p. 886–904, 2017. Citado 3 vezes nas páginas 23, 38 e 55.

FILHO, P. P. B.; ALUÍSIO, S. M.; PARDO, T. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: *9th Brazilian Symposium in Information and Human Language Technology - STIL*. Fortaleza, Brazil: [s.n.], 2013. p. 215–219. Citado na página 20.

FLEKOVA, L.; PREOTIUC-PIETRO, D.; UNGAR, L. H. Exploring stylistic variation with age and income on twitter. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. Berlin, Germany: The Association for Computer Linguistics, 2016. Citado 4 vezes nas páginas 17, 40, 53 e 55.

GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*. [S.l.]: Morgan & Claypool Publishers, 2017. ISBN 1627052984, 9781627052986. Citado 3 vezes nas páginas 20, 21 e 22.

GONZÁLEZ-GALLARDO, C. et al. Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams—Notebook for PAN at CLEF 2015. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. Toulouse, France: CEUR-WS.org, 2015. Citado 4 vezes nas páginas 17, 19, 34 e 55.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. https://deeplearningbook.com: MIT Press, 2016. ISBN 0262035618, 9780262035613. Citado na página 24.

GOPINATHAN, M.; BERG, P.-C. *A Deep Learning Ensemble Approach to Gender Identification of Tweet Authors*. Dissertação (Mestrado) — NTNU, 2017. Citado 8 vezes nas páginas 13, 14, 23, 27, 50, 53, 54 e 55.

Guimarães, R. G. et al. Age groups classification in social network using deep learning. *IEEE Access*, v. 5, p. 10805–10816, 2017. Citado 4 vezes nas páginas 17, 47, 53 e 55.

HALL, M. A. et al. The weka data mining software: an update. *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, 2009. Citado 7 vezes nas páginas 30, 31, 32, 34, 38, 40 e 42.

HARRIS, Z. Distributional structure. *Word*, v. 10, n. 23, p. 146–162, 1954. Citado na página 21.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language taks. In: *Proceesings of the Symposium in Information and Human Language Technology (STIL)*. Uberlândia, Brazil: [s.n.], 2017. Citado na página 58.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Comput.*, MIT Press, Cambridge, MA, USA, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Disponível em: ⟨http://dx.doi.org/10.1162/neco.1997.9.8.1735⟩. Citado na página 25.

HSIEH, F.; DIAS, R.; PARABONI, I. Author profiling from facebook corpora. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. [s.n.], 2018. Disponível em: ⟨http://www.lrec-conf.org/proceedings/lrec2018/summaries/354.html⟩. Citado 3 vezes nas páginas 56, 74 e 95.

HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *International AAAI Conference on Web and Social Media (ICWSM)*. [S.l.: s.n.], 2014. Citado na página 50.

ISBISTER, T.; KAATI, L.; COHEN, K. Gender classification with data independent features in multiple languages. In: *European Intelligence and Security Informatics Conference, EISIC 2017, Athens, Greece, September 11-13, 2017*. [S.l.]: IEEE Computer Society, 2017. p. 54–60. Citado 3 vezes nas páginas 19, 44 e 55.

JOULIN, A. et al. Bag of tricks for efficient text classification. In: *Conference: Conference: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. [S.l.: s.n.], 2016. Citado 2 vezes nas páginas 23 e 33.

KIM, S. M. et al. Demographic inference on twitter using recursive neural networks. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. Vancouver, Canada: [s.n.], 2017. v. 2, p. 471–477. Citado 6 vezes nas páginas 17, 27, 49, 53, 54 e 55.

KIM, Y. Convolutional neural networks for sentence classification. p. 1746–1751, 2014. Disponível em: ⟨http://aclweb.org/anthology/D/D14/D14-1181.pdf⟩. Citado 3 vezes nas páginas 14, 27 e 53.

KIM, Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. [s.n.], 2014. p. 1746–1751. Disponível em: ⟨http://aclweb.org/anthology/D/D14/D14-1181.pdf⟩. Citado 4 vezes nas páginas 27, 56, 75 e 94.

KINCAID, J. P. et al. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. [S.l.]: Chief of Naval Technical Training, Naval Air Station Memphis, 1975. (Research Branch report). Citado 2 vezes nas páginas 19 e 41.

KOSINSKI, M.; STILLWELL, D. J. *MyPersonality Project*. 2012. 110-122 p. Disponível em: ⟨http://www.mypersonality.org/wiki⟩. Citado 2 vezes nas páginas 43 e 44.

LAI, S. et al. Recurrent convolutional neural networks for text classification. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. Austin, USA: AAAI Press, 2015. v. 333, p. 2267–2273. Citado 4 vezes nas páginas 14, 27, 45 e 47.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. JMLR.org, 2014. (ICML'14), p. II–1188–II–1196. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=3044805.3045025⟩. Citado na página 58.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, MIT Press, Cambridge, MA, USA, v. 1, n. 4, p. 541–551, dez. 1989. ISSN 0899-7667. Disponível em: ⟨http://dx.doi.org/10.1162/neco.1989.1.4.541⟩. Citado na página 27.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ETMTNLP '02), p. 63–70. Disponível em: ⟨https://doi.org/10.3115/1118108.1118117⟩. Citado na página 33.

LÓPEZ-MONROY, A. P. et al. Using intra-profile information for author profiling. In: *CLEF (Working Notes)*. Sheffield, UK: CEUR-WS.org, 2014. p. 1116–1120. Citado 2 vezes nas páginas 30 e 55.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Citado 2 vezes nas páginas 19 e 41.

MARTINC, M. et al. PAN 2017: Author Profiling - Gender and Language Variety Prediction—Notebook for PAN at CLEF 2017. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*. Dublin, Ireland: CEUR-WS.org, 2017. Citado 4 vezes nas páginas 19, 23, 36 e 55.

MECHTI, S.; JAOUA, M.; BELGUITH, L. H. Author profiling using style-based features notebook for PAN at CLEF 2013. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. Valencia, Spain: CEUR-WS.org, 2013. v. 1179, p. 23–26. Citado 3 vezes nas páginas 23, 40 e 55.

MEINA, M. et al. Ensemble-based classification for author profiling using various features notebook for PAN at CLEF 2013. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. Valencia, Spain: CEUR-WS.org, 2013. p. 23–26. Citado 4 vezes nas páginas 18, 23, 31 e 55.

MEINA, M. et al. Ensemble-based classification for author profiling using various features — notebook for pan at clef 2013. In: *CLEF 2013 Evaluation Labs and Workshop*. Valencia, Spain: CEUR-WS.org, 2013. v. 1179, p. 23–26. Citado 2 vezes nas páginas 92 e 93.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. 2013. Citado 3 vezes nas páginas 22, 23 e 58.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2999792.2999959⟩. Citado 2 vezes nas páginas 36 e 47.

MUKHERJEE, A.; LIU, B. Improving gender classification of blog authors. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. p. 207–217. Citado na página 46.

NELDER, J.; MEAD, R. A simplex method for function minimization comput. *The Computer Journal*, v. 7, 01 1965. Citado na página 64.

NGUYEN, D. et al. Computational sociolinguistics: A survey. *Computational Linguistics*, MIT Press, v. 42, n. 3, p. 537–593, 2016. ISSN 0891-2017. Citado 2 vezes nas páginas 13 e 16.

NGUYEN, D. et al. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: ACL, 2014. p. 1950–1961. Citado na página 16.

NIELSEN, M. A. *Neural Networks and Deep Learning*. Determination Press, 2018. Disponível em: ⟨http://neuralnetworksanddeeplearning.com/⟩. Citado 3 vezes nas páginas 24, 25 e 27.

NIVRE, J.; FANG, C. Universal dependency evaluation. In: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Gothenburg, Sweden: Association for Computational Linguistics, 2017. p. 86–95. Citado na página 20.

ORTEGA-MENDOZA, R. M. et al. I, me, mine: The role of personal phrases in author profiling. In: *7th International Conference of the CLEF Association, CLEF 2016*. Évora, Portugal: [s.n.], 2016. v. 9822, p. 110–122. ISBN 978-3-319-44563-2. Citado na página 49.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 4 vezes nas páginas 46, 58, 59 e 61.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. *Linguistic Inquiry and Word Count*. Mahwah, NJ: Lawerence Erlbaum Associates, 2001. Citado 3 vezes nas páginas 20, 44 e 45.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. [s.n.], 2014. p. 1532–1543. Disponível em: ⟨http://www.aclweb.org/anthology/D14-1162⟩. Citado 2 vezes nas páginas 22 e 50.

POTTHAST, M. et al. Improving the reproducibility of pan's shared tasks: - plagiarism detection, author identification, and author profiling. In: *CLEF*. [S.l.]: Springer, 2014. (Lecture Notes in Computer Science, v. 8685), p. 268–299. Citado na página 62.

QIAN, Q. et al. Learning tag embeddings and tag-specific composition functions in recursive neural network. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: [s.n.], 2015. p. 1365–1374. Citado na página 49.

RAMOS, R. M. S. et al. Building a corpus for personality-dependent natural language understanding and generation. In: *11th International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: ELRA, 2018. p. 1138–1145. Citado 4 vezes nas páginas 14, 57, 67 e 95.

RANGEL, F. et al. Overview of the 3rd author profiling task at PAN 2015. In: *CLEF 2015 Evaluation Labs and Workshop, Notebook Papers, Toulouse, France, September 8-11, 2015.* [S.l.]: CEUR Workshop Proceedings, 2015. Citado 3 vezes nas páginas 17, 32 e 34.

RANGEL, F.; ROSSO, P. Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in twitter. In: CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019.* CEUR-WS.org, 2019. (CEUR Workshop Proceedings, v. 2380). Disponível em: ⟨http://ceur-ws.org/Vol-2380/paper_263.pdf⟩. Citado 4 vezes nas páginas 13, 62, 63 e 95.

RANGEL, F. et al. Overview of the 2nd author profiling task at PAN 2014. In: *CLEF 2014 Evaluation Labs and Workshop, Notebook Papers, Sheffield, UK, 2014.* [S.l.]: CEUR Workshop Proceedings, 2014. p. 1–30. Citado 2 vezes nas páginas 17 e 30.

RANGEL, F. et al. Overview of the author profiling task at PAN 2013. In: *CLEF 2013 Labs and Workshops, Notebook Papers.* [S.l.]: CEUR Workshop Proceedings, 2013. p. 352–365. Citado 8 vezes nas páginas 14, 17, 30, 31, 40, 41, 67 e 92.

RANGEL, F. et al. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter. In: *CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.* [S.l.]: CEUR Workshop Proceedings, 2017. Citado 7 vezes nas páginas 13, 14, 17, 32, 36, 42 e 48.

RANGEL, F. et al. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: *CLEF 2016 Evaluation Labs and Workshop, Notebook Papers.* [S.l.]: CEUR Workshop Proceedings, 2016. p. 750–784. Citado 3 vezes nas páginas 17, 33 e 37.

RANGEL, F. M. et al. Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in twitter. In: CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.* CEUR-WS.org, 2018. (CEUR Workshop Proceedings, v. 2125). Disponível em: ⟨http://ceur-ws.org/Vol-2125/invited_paper_15.pdf⟩. Citado 3 vezes nas páginas 14, 59 e 95.

REDDY, P.; REDDY, T. R.; VARDHAN, B. V. A survey o n authorship profiling techniques. *International Journal of Applied Engineering Research*, Research India Publications, v. 11, p. 3092 – 3102, 2016. Citado 2 vezes nas páginas 18 e 19.

REDDY, T. R.; VARDHAN, B. V.; REDDY, P. V. N-gram approach for gender prediction. In: IEEE. *Advance Computing Conference (IACC), 2017 IEEE 7th International.* Hyderabad, India, 2017. p. 860–865. Citado 5 vezes nas páginas 16, 19, 23, 35 e 55.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valletta, Malta: ELRA, 2010. p. 45–50. ⟨http://is.muni.cz/publication/884893/en⟩. Citado na página 59.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958. ISSN 0033-295X. Disponível em: ⟨http://dx.doi.org/10.1037/h0042519⟩. Citado na página 24.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. In: RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. (Ed.). Cambridge, MA, USA: MIT Press, 1986. cap. Learning Internal Representations by Error Propagation, p. 318–362. Citado na página 25.

RUSSELL, C. A.; MILLER, B. H. Profile of a terrorist. *Terrorism*, Routledge, v. 1, n. 1, p. 17–34, 1977. Disponível em: ⟨https://doi.org/10.1080/10576107708435394⟩. Citado na página 13.

SANTOS, H. D. P. dos; WOLOSZYN, V.; VIEIRA, R. BlogSet-BR: A Brazilian Portuguese Blog Corpus. In: *11th International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: ELRA, 2018. Citado 3 vezes nas páginas 14, 67 e 70.

SANTOS, W. R. dos; PARABONI, I. Moral stance recognition and polarity classification from twitter and elicited text. In: *Recents Advances in Natural Language Processing (RANLP-2019, to appear)*. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 14 e 67.

SAP, M. et al. Developing age and gender predictive lexica over social media. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*. Doha, Qatar: ACL, 2014. p. 1146–1151. Citado 5 vezes nas páginas 17, 19, 43, 53 e 55.

SCHLER, J. et al. Effects of age and gender on blogging. In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. [S.l.]: AAAI, 2006. p. 199–205. Citado 2 vezes nas páginas 14 e 67.

SIERRA, S. et al. Convolutional neural networks for author profiling. In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.* Dublin, Ireland: CEUR-WS.org, 2017. Citado 6 vezes nas páginas 14, 17, 23, 48, 53 e 55.

SOCHER, R. et al. Parsing natural scenes and natural language with recursive neural networks. In: GETOOR, L.; SCHEFFER, T. (Ed.). *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. [S.l.]: Omnipress, 2011. p. 129–136. Citado na página 49.

SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, JMLR.org, v. 15, n. 1, p. 1929–1958, jan. 2014. ISSN 1532-4435. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2627435.2670313⟩. Citado na página 86.

TAKAHASHI, T. et al. Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*. CEUR-WS.org, 2018. (CEUR Workshop Proceedings, v. 2125). Disponível em: ⟨http://ceur-ws.org/Vol-2125/paper_83.pdf⟩. Citado 4 vezes nas páginas 14, 56, 63 e 75.

VOLLENBROEK, M. B. op et al. Gronup: Groningen user profiling. In: *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*. [S.l.]: CEUR-WS.org, 2016. p. 846–857. Citado 5 vezes nas páginas 19, 23, 33, 37 e 55.

WANG, Y. et al. Attention-based LSTM for aspect-level sentiment classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2016. p. 606–615. Citado 4 vezes nas páginas 26, 56, 75 e 94.

WEREN, E. R. D. et al. Examining multiple features for author profiling. *Journal of Information and Data Management (JIDM)*, v. 5, n. 3, p. 266–279, 2014. Citado 3 vezes nas páginas 19, 41 e 55.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: *Advances in neural information processing systems*. Montreal, Canada: [s.n.], 2015. p. 649–657. Citado 2 vezes nas páginas 14 e 27.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. Cambridge, MA, USA: MIT Press, 2015. (NIPS'15), p. 649–657. Disponível em:

⟨http://dl.acm.org/citation.cfm?id=2969239.2969312⟩. Citado 4 vezes nas páginas 56, 75, 76 e 94.

ZHANG, Y.; WALLACE, B. C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In: *IJCNLP(1)*. [S.l.]: Asian Federation of Natural Language Processing, 2017. p. 253–263. Citado na página 28.

ZHU, X.; KIRITCHENKO, S.; MOHAMMAD, S. NRC-canada-2014: Recent improvements in the sentiment analysis of tweets. In: *Proceedings of the 8th International Workshop on Semantic Evaluation Exercises (SemEval-2014), Dublin, Ireland, August, 2014*. [S.l.: s.n.], 2014. Citado na página 41.