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

ALEX GWO JEN LAN

Classificação computacional de fundamentos morais a partir de texto

São Paulo
2022

ALEX GWO JEN LAN

Classificação computacional de fundamentos morais a partir de texto

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo como parte dos requisitos 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 28 de março de 2022. 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

2022

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Lan, Alex Gwo Jen
Classificação computacional de fundamentos morais a partir de texto / Alex Gwo Jen Lan; orientador, Ivandre Paraboni. -- São Paulo, 2022.
88 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Fundamentos morais. 2. Classificação de texto. 3. Análise de sentimentos. 4. Caracterização autoral. I. Paraboni, Ivandre, orient. II. Título.

Dissertação de autoria de Alex Gwo Jen Lan, sob o título **"Classificação computacional de fundamentos morais a partir de texto"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, como parte dos requisitos 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 28 de março de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Presidente

Prof. Dr. Evandro Eduardo Seron Ruiz

Universidade de São Paulo

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Prof. Dr. Thiago Alexandre Salgueiro Pardo

Universidade de São Paulo

Instituto de Ciências, Matemáticas e de Computação

Agradecimentos

Agradeço sinceramente a todas as pessoas que me apoiaram e me motivaram na realização desse mestrado. À minha família pelo apoio incondicional e compreensão nos momentos de ausência para me dedicar ao projeto. Aos professores do Programa de Pós-graduação em Sistemas de Informação (PPgSI) da EACH por todos os ensinamentos e pela inspiração em seguir na academia, em especial à Profa. Dra. Sarajane Peres. E acima de tudo, ao meu orientador, Prof. Dr. Ivandré Paraboni, pela paciência, motivação e por todo o acompanhamento ao longo dos anos, desde a época da iniciação científica na graduação. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 88887.475828/2020-00.

Resumo

LAN, Alex Gwo Jen. **Classificação computacional de fundamentos morais a partir de texto**. 2022. 88 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

A identificação de valores morais em textos e discursos humanos contribui essencialmente para a compreensão de conflitos sociais motivados pelas diferenças de moralidade, além de comportamentos e posições ideológicas individuais. Em vista disso, muitas são as suas aplicações para a modelagem de problemas e eventos sociais, envolvendo análise de debates políticos, identificação de notícias falsas e o desencadeamento de eventos como protestos, campanhas eleitorais, por exemplo. O presente trabalho apresenta um estudo de classificação de categorias morais a partir de textos pautado sobre a Teoria dos Fundamentos Morais (TFM) com a utilização de métodos supervisionados de aprendizado de máquina. Estas categorias consistem em Cuidado, Justiça, Lealdade, Autoridade e Pureza. A tarefa é definida de duas formas sob a perspectiva do Processamento de Língua Natural (PLN). A primeira delas trata da classificação de fundamentos morais impessoais (CFMI), que é abordada de maneira similar às tarefas de análise de sentimentos, no sentido de que os fundamentos são referentes apenas ao significado expresso no texto. Por outro lado, a tarefa de classificação de fundamentos morais pessoais (CFMP), que é essencialmente inexplorada na literatura, define-se como uma instância de caracterização autoral, ou seja, considera a moralidade do autor do texto analisado, permitindo assim a captura de informação de classe não necessariamente explícita. Os trabalhos existentes são baseados em formas de representação textual mais tradicionais como *Bag-Of-Words* e *word embeddings* estáticos. Como forma de avançar o estado-da-arte nesses dois tipos de problema, este estudo propõe o desenvolvimento de modelos baseados em métodos de *embeddings* sensíveis ao contexto para as tarefas de CFMI e CFMP. De forma específica, foram conduzidos experimentos com estas duas tarefas nos idiomas em inglês, para CFMI, e português brasileiro, para CFMP, utilizando modelos como ELMo e BERT. Os resultados sugerem a eficácia no uso desses *embeddings* sensíveis ao contexto em CFMI e o potencial dos modelos de CFMP baseados em métodos como regressão logística com n-gramas de caracteres. Com isso, deixam-se oportunidades de futuros estudos na área, especialmente para PLN em português brasileiro.

Palavras-chaves: Fundamentos morais. Classificação de texto. Análise de Sentimentos. Caracterização Autoral.

Abstract

LAN, Alex Gwo Jen. **Computational classification of moral foundations from text**. 2022. 88 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2022.

The identification of moral values in human texts and speeches essentially contributes to the understanding of social conflicts motivated by differences in morality, in addition to individual behaviours and ideological positions. For this reason, there are many applications based on moral identification for modelling social problems and events, involving analysis of political debates, fake news identification and prediction of events such as protests, election campaigns, for example. This work presents a study of moral categories classification from text based on Moral Foundations Theory using machine learning supervised methods. These categories consist of Care, Fairness, Loyalty, Authority e Purity. The task is defined in two ways from the perspective of Natural Language Processing (NLP). The first one deals with the impersonal moral foundations classification (IMFC), which is approached in a similar fashion to the tasks of sentiment analysis, in the way that foundations refer only to the meaning expressed in the text. On the other hand, the personal moral foundations classification (PMFC) task, which is essentially unexplored in the literature, defines itself as an instance of author profiling, that is, it considers the morality of the author of the analysed text, thus allowing the capture of class information that is not necessarily explicit. Existing works are based on more traditional methods of textual representation such as Bag-Of-Words and static word embeddings. As a way to advance the state-of-the-art in these two types of problems, this study proposes the development of models based on contextual-sensitive embeddings methods for IMFC and PMFC. Specifically, experiments were conducted with these two tasks in English, for IMFC, and Brazilian Portuguese, for PMFC, using models such as ELMo and BERT. The results suggest the effectiveness of using these contextual-sensitive embeddings in IMFC and the potential of PMFC models based on methods such as logistic regression with character n-grams. This provides opportunities for future studies, especially for NLP in Brazilian Portuguese.

Keywords: Moral Foundations. Text Classification. Sentiment Analysis. Author Profiling.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

BERT	*Bidirectional encoder representations from transformers*
BoW	*Bag-Of-Words*
CFMI	Classificação de fundamentos morais impessoais
CFMP	Classificação de fundamentos morais pessoais
DFM	Dicionário de fundamentos morais
ELMo	*Embeddings from Language Models*
FFNN	*Feed Forward Neural Network*
GRU	*Gated recurrent units*
HLMRF	*Hinge-loss Markov random field*
LSTM	*Long short term memory*
MFTC	*Moral foundations Twitter corpus*
PLN	Processamento de língua natural
QFM	Questionário de fundamentos morais
RNA	Redes neurais artificiais
RNN	Redes neurais recorrentes
TFM	Teoria dos fundamentos morais

Sumário

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

1 Introdução

A moralidade é um sistema de valores e princípios que determinam o que é admitido ou não dentro de um grupo social, além de ser um fator que molda consideravelmente as decisões e atitudes das pessoas. Estudos acerca desse conceito na ciência política, psicologia, e principalmente na computação, são apoiados pela Teoria dos Fundamentos Morais (TFM) (GRAHAM *et al.*, 2013). Segundo essa teoria, a perspectiva moral humana é caracterizada por ser inata aos grupos sociais (nativismo), passível de transformação em seu meio (aprendizagem cultural), guiada por processos inconscientes e automáticos (primazia a intuição) e por sua pluralidade. Este último, define a moralidade em termos de cinco fundamentos morais: Cuidado, Justiça, Lealdade, Autoridade e Pureza.

Cada um dos fundamentos morais pode ser expresso de forma subjacente em discursos humanos, por termos que o apoiem (virtude), ou que o violam (vício). Para fins de ilustração, dois *tweets* relacionados com #AllLivesMatter são extraídos do córpus *Moral Foundations Twitter* (HOOVER *et al.*, 2020), adaptados para o português brasileiro, e apresentados abaixo. O primeiro exemplo refere-se a um texto com apoio à categoria de Autoridade, enquanto que o segundo, a um texto com a violação do mesmo fundamento.

> *Exemplo 1: Ensine suas crianças a obedecer a lei e a respeitar os outros. A mudança começa em casa. #Ferguson #BlackLivesMatter #AllLivesMatter*

> *Exemplo 2: Cansado desses policiais abusando de seus poderes e autoridade! Vocês estão aqui para PROTEGER as pessoas e não causar dano a elas. #AllLivesMatter*

A identificação de valores morais em discursos humanos desempenha um papel importante na compreensão de conflitos sociais motivados pelas diferenças de moralidade, além de comportamentos (CURRY; MULLINS; WHITEHOUSE, 2019; KALIMERI *et al.*, 2019) e posições ideológicas frente a uma diversidade de assuntos (KOLEVA *et al.*, 2012), e.g., doações (WINTERICH; ZHANG; MITTAL, 2012; HOOVER *et al.*, 2018), pobreza (LOW; WUI, 2016), vacinação (AMIN *et al.*, 2017), mudanças climáticas (WOLSKO; ARICEAGA; SEIDEN, 2016) e entre outros. Ademais, reconhecer esses valores é importante para avaliar perfis políticos, uma vez que indivíduos de divergentes orientações (e.g., liberal e conservador) atendem diferentes intuições morais (GRAHAM; HAIDT; NOSEK, 2009; SYLWESTER; PURVER, 2015; FULGONI *et al.*, 2016).

Do ponto de vista computacional, a necessidade de inferir fundamentos morais a partir de grandes volumes de texto de forma automática tem motivado diversos estudos de Processamento de Língua Natural (PLN) e áreas afins (TEERNSTRA *et al.*, 2016; GARTEN *et al.*, 2018; LIN *et al.*, 2018). Estudos deste tipo tipicamente fazem uso de métodos de aprendizado de máquina supervisionado para explorar a relação entre o conteúdo textual e seus inerentes valores morais. O presente trabalho é também um representante deste tipo de pesquisa.

O problema de classificação de fundamentos morais a partir de texto é tradicionalmente representado na literatura como a tarefa de inferir categorias da TFM a partir de um texto de cunho moral, como uma opinião sobre uma questão social polêmica (e.g., postagens de rede social e discursos políticos). Esta tarefa será aqui denominada *classificação de fundamentos morais impessoais*, ou CFMI. O termo "impessoal" refere-se ao fato de que essa tarefa avalia o texto de forma independente da moralidade do indivíduo que o escreveu, ou seja, considerando apenas a informação presente no texto.

Os estudos de CFMI como Teernstra *et al.* (2016) e Lin *et al.* (2018) são uma aplicação direta de métodos de classificação impessoais, tipicamente fazendo uso de textos rotulados com categorias da TFM expressas no texto para determinar se a ideia inerente é, por exemplo, de cunho mais liberal ou conservador. Assim, desde que corretamente rotulados, textos semanticamente próximos (e.g., duas opiniões semelhantes sobre um determinado assunto) seriam associados às mesmas categorias da TFM, e possivelmente classificados da mesma forma.

Por outro lado, uma segunda definição possível para o problema de classificação de fundamentos morais a partir de texto, e que é inédita na literatura, seria a inferência de categorias da TFM associadas ao *indivíduo* que escreveu um determinado texto. Esta tarefa, aqui denominada *classificação de fundamentos morais pessoais*, ou CFMP, pode ser vista como uma instância do problema de caracterização autoral (RANGEL; ROSSO, 2019) comumente utilizada na inferência de variáveis sociais do autor de um texto, como gênero (TAKAHASHI *et al.*, 2018), faixa etária (RANGEL *et al.*, 2020), traços de personalidade (SANTOS; RAMOS; PARABONI, 2019) e outros.

Um estudo de CFMP faria uso de textos rotulados com categorias da TFM que expressam a moralidade do indivíduo, obtidas por meio de um instrumento adequado (como o questionário de fundamentos morais em Silvino *et al.* (2016)) e, diferentemente do problema de CFMI tradicional, constituiria um problema de classificação indireta em que

a informação da classe a ser apreendida não necessariamente está explícita no texto. Em outras palavras, uma mesma opinião sobre um determinado assunto poderia ser rotulada de forma distinta a depender de quem a escreveu.

Os problemas de CFMI e CFMP, aqui tratados coletivamente na classificação de fundamentos morais a partir de texto, serão o foco deste trabalho. Os modelos de CFMI encontrados na literatura (TEERNSTRA *et al.*, 2016; LIN *et al.*, 2018; NOKHIZ; LI, 2017; GARTEN *et al.*, 2018) são geralmente baseados em representações textuais tradicionais como o modelo *Bag-Of-Words*, ou em representação distribuída de palavras do tipo estático como produzida por métodos *Word2Vec* (MIKOLOV *et al.*, 2013a). Neste tipo de tarefa, modelos mais recentes baseados em *embeddings* contextuais, apesar dos resultados superiores em diversos trabalhos (JOSHI *et al.*, 2019; PECAR; SIMKO; BIELIKOVÁ, 2019; ZAMPIERI *et al.*, 2019), permanecem quase inexplorados. No caso do problema de CFMP, não encontramos nenhum exemplo deste tipo de estudo na literatura de PLN.

Estas observações sugerem a oportunidade de estudo e desenvolvimento de modelos de CFMI e CFMP utilizando modelos de língua pré-treinados do tipo ELMo (PETERS *et al.*, 2018) ou BERT (DEVLIN *et al.*, 2019). Esta dissertação apresenta os resultados de uma pesquisa em nível de mestrado acadêmico abordando esta questão.

1.1 Hipóteses

> H1: O uso de modelos de língua pré-treinados permite classificar fundamentos morais expressos em textos - ou CFMI - com resultados superiores aos de modelos de classificação tradicionais.
>
> H2: O uso de modelos de língua pré-treinados permite classificar fundamentos morais de um indivíduo com base em textos de sua autoria - ou CFMP - com resultados superiores aos de modelos de classificação tradicionais.

Essas hipóteses serão testadas comparando-se os modelos propostos do tipo ELMo e BERT com os sistemas de *baseline* definidos a partir de estudos já existentes quando possível ou modelos baseados em representações textuais tradicionais. Na avaliação, será utilizada uma das tradicionais métricas em tarefas de aprendizado de máquina, a F1 macro.

1.2 Objetivo

O presente trabalho propõe o desenvolvimento de modelos de classificação de fundamentos morais a partir de texto (CFMI e CFMP) baseados em modelos de língua pré-treinados de representação contextual, de modo a obter resultados superiores aos de classificadores de texto tradicionais.

De forma mais específica, a tarefa de classificação de fundamentos morais expressos no texto, CFMI, será abordada com o uso de um córpus anotado em língua inglesa, o *Moral Foundations Twitter Corpus* (MFTC) (HOOVER *et al.*, 2020), e a classificação indireta de fundamentos morais de um indivíduo com base em textos de sua autoria, CFMP, fará uso de um córpus em português brasileiro, o BRmoral (PAVAN *et al.*, 2020). A escolha em utilizar estas línguas nas respectivas tarefas foi devida à disponibilidade do conjuntos de dados anotados na literatura durante o desenvolvimento do presente projeto de pesquisa.

1.3 Organização do documento

Os próximos capítulos apresentam os conceitos fundamentais para a compreensão do presente estudo (Capítulo 2), as discussões dos trabalhos correlatos da literatura (Capítulo 3), os materiais e métodos (Capítulo 4), os resultados dos experimentos realizados (Capítulo 5) e, por fim, a conclusão deste trabalho (Capítulo 6).

2 Conceitos fundamentais

Este capítulo descreve os conceitos fundamentais associados ao escopo desta pesquisa. Na primeira seção, são apresentadas as definições relacionadas à Teoria dos Fundamentos Morais (TFM), assim como ferramentas para identificar moralidade em texto. Na seção seguinte, são introduzidos conjuntos de dados anotados com informação de moralidade que estão disponíveis na literatura. E por fim, nas últimas seções, são discutidas as técnicas de aprendizado de máquina e de representação textual usadas para abordar as tarefas de classificação de fundamentos morais a partir de textos.

2.1 Teoria dos fundamentos morais (TFM)

A Teoria dos Fundamentos Morais (TFM) foi elaborada por psicólogos sociais para explicar a moralidade e suas diferenças entre as diversas culturas, assim como as semelhanças compartilhadas entre elas. A teoria é construída a partir de quatro pilares: o nativismo, o aprendizado cultural, o intuicionismo e o pluralismo (GRAHAM *et al.*, 2013).

O nativismo discorre sobre o caráter inato da moralidade nos indivíduos. Segundo a teoria, existe um rascunho inicial escrito pelos genes na mente moral. Este rascunho foi organizado ao longo da história evolutiva da espécie humana em resposta às pressões adaptativas. Os autores utilizam como exemplo a maior facilidade em ensinar as crianças a desejarem por vingança quando se sentem injustiçadas do que perdoarem os seus inimigos.

Apesar da sua presença intrínseca na psique humana, a moralidade é suscetível à influência do aprendizado cultural. Desse modo, a variedade de contextos sociais proporciona essa variação da moralidade entre as diversas culturas. Esse primeiro rascunho seria editado de maneiras distintas por normas sociais regidas pela predominância de algumas das ordens morais.

De acordo com o intuicionismo, as avaliações morais das pessoas são processos rápidos, automáticos e com ausência de esforço. Esse mecanismo de julgamento moral, descrito pelo modelo intuicionista social, ocorre em duas fases: o contato inicial com o objeto e a determinação do mesmo como algo moralmente positivo ou negativo, e em seguida, a justificativa para essa avaliação.

No aspecto do pluralismo, como reação aos desafios adaptativos enfrentados pelos ancestrais ao longo do tempo, foram moldadas estruturas mentais inatas, ao qual denominou-se "fundamentos morais". O termo "fundamento" foi empregado para transmitir a ideia de que estas estruturas não se encerram como definições fechadas ou "edificações prontas", mas constituem o alicerce moral das mais variadas culturas. No estado atual da teoria, limitou-se a cinco fundamentos morais, porém, acredita-se que possam existir ainda mais categorias. Com detalhes, os fundamentos são descritos abaixo, de acordo com as definições de Graham *et al.* (2013):

- O Cuidado é relacionado com a compreensão pelo sofrimento do outro com sentimentos de proteção, nutrição ou cuidado. Este fundamento está presente entre os mamíferos, especialmente na maternidade. Esta estrutura se torna mais ativa nas mães para atender as necessidades de seus filhotes.
- A Justiça está associada aos conceitos de igualdade, direitos, altruísmo recíproco e cooperação. Para animais sociáveis, os mais vantajosos em questão de sobrevivência foram aqueles capazes de perceber situações de trapaça e reagir com o intuito de alcançar uma condição mais justa.
- A Lealdade refere-se ao patriotismo, dedicação ou o auto-sacrifício a um grupo. No passado da linha evolutiva dos homens, existiram competições entre grupos de primatas, e sobressaíam vencedores aqueles que conseguiam formar coalizões coesivas. Como exemplo, esse fundamento é observado em fãs de esporte e de marcas.
- A Autoridade descreve a hierarquia, o respeito às tradições e a obediência dentro de um grupo. Este fundamento está presente em interações com chefes e líderes, em instituições como tribunais da lei e departamentos da polícia.
- A Pureza é descrita com o sentimento de desgosto, e está relacionada a um sistema de imunidade comportamental. Ela é subjacente às noções religiosas de manter uma condição de vida mais elevada. Na história evolutiva, os hominídeos ancestrais se expuseram a riscos com patógenos, e para se afastar deles, eles desenvolveram uma dieta mais onívora com mais carne. No contexto atual, esse fundamento também é empregado para suportar algumas reações direcionadas a grupos sociais (e.g., imigrantes e homossexuais).

A TFM é aplicada em estudos que analisam o conteúdo textual para recuperação de informações à respeito do discurso moral, contribuindo para a modelagem de uma série

de problemas sociais (KOLEVA *et al.*, 2012). Para isso, os autores introduziram duas ferramentas para mensurar ou verificar a presença de fundamentos morais no discurso, que são o Questionário de Fundamentos Morais e o Dicionário de Fundamentos Morais.

2.1.1 Questionário de fundamentos morais (QFM)

A partir da necessidade de mensurar moralidade presente no discurso humano, em Graham *et al.* (2011) é desenvolvido o Questionário de Fundamentos Morais (QFM). Trata-se de uma abordagem amplamente utilizada em laboratórios e via configurações remotas para trabalhos no âmbito da psicologia moral (GRAHAM *et al.*, 2013). Dada a sua importância, o questionário recebeu uma adaptação para o português brasileiro (SILVINO *et al.*, 2016).

O QFM estima para cada indivíduo pontuações relacionadas aos cinco fundamentos da TFM. Este questionário é constituído por 32 itens de múltipla escolha cuja respostas estão definidas dentro de uma escala de cinco pontos, que vão de "discordo totalmente" a "concordo totalmente". Além disso, o QFM está dividido em duas partes principais: questões relacionadas à relevância de certas variáveis para o julgamento moral (e.g., a importância do sofrimento emocional de uma pessoa), e o grau de concordância sobre algum assunto moral (e.g., toda criança precisa aprender a respeitar autoridades).

Apesar de prover informações importantes sobre as pontuações de moralidade, além das adicionais (e.g., sociodemográficos e pessoais), o questionário possui desvantagens, como estar associado a uma limitada escalabilidade e altos custos de aplicação. Como alternativa, alguns estudos optaram por abordagens baseadas no dicionário de fundamentos morais para analisar córpus.

2.1.2 Dicionário de fundamentos morais (DFM)

O Dicionário de Fundamentos Morais (DFM) é um recurso psicolinguístico, introduzido num dos estudos de Graham, Haidt e Nosek (2009), que contém lemas e radicais em inglês associados às categorias da TFM. Essas palavras oferecem evidência para o fundamento moral correspondente num texto. Em sua versão mais recente, o dicionário é constituído por 324 termos.

Os termos do DFM estão distribuídos entre os cinco fundamentos, e uma categoria adicional, a "miscelânea". Nesta última, estão contidas palavras com relevância moral e que ainda não se adequam a nenhuma das outras categorias. Cada categoria possui uma subdivisão binária entre virtude e vício, sendo que o primeiro refere-se aos termos que apoiam o dado fundamento (e.g., *safe* para Cuidado) e o segundo refere-se àqueles que o violam (e.g., *kill* para Cuidado).

A aplicação do DFM para a tarefa de classificação a partir de texto consiste na análise do número de ocorrências das palavras do dicionário que se encontram dentro de um documento de entrada (e.g., *tweet*, discurso político e etc.) (GRAHAM; HAIDT; NOSEK, 2009). Como exemplo, considera-se o documento "devemos respeitar e cuidar da nossa cidade", que seria classificado com os fundamentos Autoridade e Cuidado, por conter termos do DFM associados a estes dois fundamentos (i.e., respeitar para Autoridade e cuidar para Cuidado). Mais exemplos de trabalhos deste tipo são descritos na seção de abordagens baseadas no DFM da revisão bibliográfica (seção 3.1).

O DFM é um recurso de grande serventia aos estudos que se apoiam na TFM, e é notável a necessidade e o interesse por alguns dos pesquisadores da área em adaptá-lo para o estudo da moralidade em outras culturas. Existem versões do DFM nos idiomas japonês, o denominado J-DFM (MATSUO *et al.*, 2019), em madarim (WANG; LIU; YU, 2020) e, inclusive, em português brasileiro (CARVALHO *et al.*, 2020).

Em Rezapour, Shah e Diesner (2019) demonstrou-se também que a expansão do DFM pode melhorar os resultados dos modelos que a utilizam. Para isso, os autores deste trabalho apresentaram uma estratégia de expansão do DFM com o incremento de termos da *Wordnet* e desambiguação com *Part-Of-Speech*.

Nos estudos de Hopp *et al.* (2020), é relatado o desenvolvimento de outra expansão, o eMFD (*extended Moral Foundations Dictionary*). Esta versão foi construída a partir do trabalho de 824 anotadores humanos sobre 2.995 artigos de notícias em inglês de grandes jornais estadunidenses (entre novembro de 2016 a janeiro de 2017). Diferente do DFM original, cada termo presente no eMFD é associado a um vetor de probabilidade de cinco dimensões, na qual cada posição é relacionada a um dos fundamentos, e a um outro vetor de probabilidades de duas dimensões (vício e virtude).

Finalmente, o estudo em Araque, Gatti e Kalimeri (2020) também introduziu uma versão expandida, a *MoralStrength*, que é construída a partir de *synsets* da *Wordnet*. Este dicionário oferece aproximadamente três vezes mais lemas em comparação com o

DFM. Além disso, para cada termo é associado um valor numérico de valência moral, determinando a força para um fundamento específico.

2.2 Córpus rotulados com informações de fundamentos morais

Como conjunto de dados para a análise textual da abordagem proposta, serão considerados no presente trabalho os córpus *Moral Foundations Twitter Corpus* (HOOVER *et al.*, 2020) e o BRmoral (PAVAN *et al.*, 2020). Estes dois conjuntos de dados foram selecionados por representarem respectivamente os maiores córpus rotulados com informações de fundamentos morais em inglês e em português brasileiro, no momento do desenvolvimento deste trabalho.

2.2.1 O córpus *Moral Foundations Twitter* (MFTC)

O córpus *Moral Foundations Twitter* (MFTC) (HOOVER *et al.*, 2020) é uma coleção de 35.108 *tweets* em língua inglesa rotulados para as 10 categorias morais da TFM, i.e., considerando o vício e a virtude para cada um dos cinco fundamentos. Neste córpus, a anotação foi realizada em nível de texto (*tweet*), ou seja, sem considerar a identidade do indivíduo que o produziu. Dessa forma as mensagens contendo opiniões semelhantes devem estar rotuladas de forma similar. Esse córpus é adequado ao desenvolvimento de modelos de caráter "impessoal"(CFMI) a serem discutidos no capítulo 4.

O córpus MFTC foi elaborado com o intuito de contribuir para pesquisas que integram psicologia e processamento de língua natural. A seleção dos *tweets* que compõem o córpus considerou como critério a relevância com os problemas da atualidade em ciências sociais e a variedade de preocupações morais abrangidas. O MFTC é organizado em sete domínios de discurso:

- 2016 *Presidential Election*. Contém os *tweets* postados durante o período de eleição presidencial dos EUA de 2016. Os dados foram selecionados a partir de seguidores de @HillaryClinton, @realDonaldTrump, @NYTimes, @washingtonpost e @WSJ.
- *All Lives Matter*. Contém os *tweets* postados entre 2015 a 2016 que apresentam as *hashtags* #BluesLivesMatter e #AllLivesMatter.

- *Baltimore Protests*. Contém os *tweets* postados entre 5 de agosto a 4 de dezembro de 2015. Estão relacionados com os protestos à morte de Freddie Gray.
- *Black Lives Matter*. Contém os *tweets* postados entre 2015 a 2016, que apresentam as *hashtags* #BLM e #BlackLivesMatter.
- *Davidson*. Os *tweets* pertencem à Davidson *et al.* (2017). Referem-se à postagens com discurso de ódio e conteúdo ofensivo.
- *Hurricane Sandy*. Contém os *tweets* postados momentos próximos da ocorrência do furacão *Sandy* (entre 16 de outubro a 5 de novembro de 2012).
- *MeToo*. Contém as postagens de 200 indivíduos acerca do movimento #MeToo.

O conjunto de dados foi fornecido na íntegra mediante a solicitação direta aos autores do estudo. A Tabela 1 descreve a relação entre os domínios e fundamentos morais no MFTC. Para cada fundamento, as categorias de vício e virtude estão representadas respectivamente, pelos símbolos (-) e (+). Além disso, a categoria "Não-moral", refere-se aos *tweets* que não carregam o conceito de moralidade.

Tabela 1 – Relação entre domínios e fundamentos morais no MFTC

Fundamento	ALM	Balt.	BLM	Election	Davidson	Sandy	MeToo
Autoridade (-)	392	1700	701	484	142	76	2285
Autoridade (+)	620	666	606	527	1563	1196	1454
Justiça (-)	1220	1423	1558	1053	401	1072	1466
Justiça (+)	1235	700	1349	1037	194	1044	1022
Cuidado (-)	1777	1040	2094	1161	477	562	1074
Cuidado (+)	1294	610	1142	1048	103	1585	675
Lealdade (-)	409	1612	569	481	301	1572	1344
Lealdade (+)	788	1142	918	791	310	647	894
Pureza (-)	443	267	630	490	574	1515	2208
Pureza (+)	322	228	509	1171	56	396	521
Não-moral	3037	4079	3753	4372	4117	1421	2391
Total	4424	5593	5257	5358	4994	4591	4891

Fonte: Adaptada de Hoover *et al.* (2020)

2.2.2 O córpus BRmoral

O córpus BRmoral (PAVAN *et al.*, 2020) é uma coleção de textos rotulados com informação da TFM no nível do indivíduo provenientes de respostas fornecidas pelos participantes da coleta de dados para o QFM. Neste córpus, duas opiniões semelhantes (e.g.,

à favor de cotas raciais) podem ter sido rotuladas com fundamentos morais completamente distintos a depender de como o QFM foi respondido por cada participante e portanto a tarefa de classificar fundamentos morais de indivíduos a partir de texto representa uma estimativa geral de característica demográfica (i.e., moralidade) do autor do texto, e não do significado exato do texto. Assim, o BRmoral é indicado para o desenvolvimento de modelos de caráter "pessoal"(CFMP) a serem discutidos no capítulo 4.

O intuito do BRmoral é oferecer suporte para tarefas de processamento de línguas relacionadas à caracterização autoral, ao reconhecimento de posicionamentos e à classificação de fundamentos morais. O córpus contém posicionamentos de indivíduos à respeito de oito tópicos de natureza moral (casamento *gay*, cotas raciais, isenção de taxas para as igrejas, legalização das drogas, lei do aborto, porte de armas, pena de morte e a redução da maioridade penal) em português brasileiro. Abaixo, são apresentados dois exemplos de posicionamentos, o primeiro favorável e o segundo contrário à legalização do aborto:

> *Exemplo 1 (favorável): O aborto deveria ser legalizado, pois evitaria que crianças sem a menor condição de serem mantidas por seus pais sofressem em uma sociedade sem nenhum tipo de suporte financeiro, moral e educacional. Além disso, cada pessoa é dono de seu corpo e deve fazer aquilo que achar melhor com o mesmo.*

> *Exemplo 2 (contrário): O aborto consiste em ação contra a vida de indivíduo indefeso. Isto se chama assassinato. É mais latente esta interpretação nos casos onde o aborto é defendido como simples "opção" e direito para as mães/pais de decidirem sobre a vida do filho, mesmo em casos indesejados, como estupro.*

Os dados foram coletados em colaboração com outros pesquisadores do grupo no qual o presente trabalho está inserido a partir das respostas fornecidas anonimamente por voluntários a uma pesquisa *online*, que é embasada nas questões do QFM em português. Os 510 participantes são maiores de 18 anos, falantes nativos do português brasileiro, na média de 29,4 anos e majoritariamente masculina (65,8%). Ao total, o conjunto contém 4.080 textos de posicionamento (510 participantes * 8 tópicos).

Os textos do córpus estão disponibilizados com rótulos numéricos de cada uma das cinco dimensões morais da TFM, calculados a partir das respostas dos questionários, segundo o método em Graham *et al.* (2011). Além disso, os textos também estão disponibilizados com os rótulos discretos "*Low*","*Avg*" e "*High*". As instâncias com rótulos numéricos abaixo da média menos o desvio padrão de 0,5 pertencem à classe "*Low*", enquanto que aquelas que apresentam pontuações acima da média mais o desvio padrão de

0,5, à classe "*High*", e as demais, à classe "*Avg*". Essa distribuição é conveniente do ponto de vista de métodos de aprendizado de máquina a serem adotados no presente trabalho, e está ilustrada na Tabela 2.

Tabela 2 – Distribuição de classes para fundamentos morais no BRMoral

Classe	Low	Avg	High
Cuidado	157	181	159
Justiça	121	221	155
Lealdade	155	195	147
Autoridade	152	196	149
Pureza	177	183	137

Fonte: Adaptada de Pavan *et al.* (2020)

2.3 Métodos de aprendizado de máquina

Segundo Teernstra *et al.* (2016), os trabalhos mais recentes de classificação de fundamentos morais empregam métodos de aprendizado de máquina, e tendem a depender cada vez menos do DFM. A seguir, são apresentados os conceitos fundamentais sobre a regressão logística, as redes neurais artificiais (RNA), as redes neurais recorrentes (RNNs) e as redes *encoder-decoder*.

2.3.1 Regressão logística

A regressão logística é um classificador probabilístico de aprendizado de máquina supervisionado, i.e., para realizar as suas predições, este modelo recebe como entrada um conjunto de características X e de rótulos Y. O desempenho deste classificador pode ser pensado como uma busca da probabilidade de ocorrência de uma das classes de Y dada uma instância de X. Por exemplo, é viável considerar uma aplicação em que $P(y = 1 \mid x)$ é a probabilidade de uma instância apresentar um sentimento positivo, e $P(y = 0 \mid x)$, um sentimento negativo, para um determinado *tweet* x (JURAFSKY; MARTIN, 2019).

O processamento da regressão logística considera vetores de características X, de rótulos Y e de pesos W, além de um termo de viés b. Cada peso em W, w_i, é um número real que corresponde à importância das respectivas instâncias x_i. A partir de algum desses valores de entrada, é definido o termo *logit*, mostrado na fórmula a seguir.

$$z = (\sum_{i=1}^{n} w_i x_i) + b = WX + b$$

Como mostrado na fórmula, o valor do *logit* z é descrito como o produto escalar entre os vetores W e X, somado ao termo b. Após o cálculo, *logit* z é passado como entrada para uma função sigmóide, também denominada logística. A função sigmóide, descrita na próxima fórmula, produz um valor contínuo compreendido dentro do intervalo entre zero e um, o que pode ser usado como uma probabilidade. Com um limiar de decisão determinado (geralmente, o valor de 0,5), é definida uma estimativa de predição $\hat{y}$. Valores de sigmóide de z superiores a esse limiar são atribuídos a uma classe $\hat{y} = 1$, caso contrário, $\hat{y} = 0$, por exemplo.

$$\hat{y} = \sigma(z) = \frac{1}{(1 + e^{-z})}$$

A regressão logística apresenta duas etapas. A primeira delas é a de treinamento, aos quais os hiperparâmetros do modelo (o vetor de pesos W e o viés b) são aprendidos com os algoritmos de custo (entropia cruzada) e de otimização (gradiente de descida estocástico). Na segunda etapa, a de teste, $P(y \mid x)$ é computada para um dado x, retornando o rótulo ($y = 1$ ou $y = 0$) com maior probabilidade.

2.3.2 Redes neurais artificiais (RNA)

As redes neurais artificiais (RNAs) são uma família de classificadores construídos sobre pequenas unidades computacionais, os neurônios. Cada uma dessas pequenas estruturas recebe um vetor de valores como entrada e produz uma única saída. Essas unidades são dispostas em camadas organizadas ao longo da arquitetura da rede (JURAFSKY; MARTIN, 2019).

Um neurônio opera transformações lineares sobre as suas entradas, calculando a soma ponderada delas, com a adição de um termo de viés ($z = Wx + b$). Sobre esse valor resultante, é aplicada uma função de ativação (e.g., sigmóide, *softmax*, ReLU, tanh e etc.), conferindo-lhes o aspecto da não-linearidade, que possibilita a solução de problemas mais complexos (GOODFELLOW; BENGIO; COURVILLE, 2016)

As primeiras variações de RNAs são as *Feed Forward Neural Networks* (FFNN), caracterizadas pela arquitetura multicamada e conexões não cíclicas entre os neurônios. Os sinais são processados para frente de um neurônio a outro de camada distinta. Essa informação flui adiante pelos três tipos de camadas presentes na arquitetura: a entrada, na qual um vetor de valores é recebido, a escondida, onde ocorre as principais computações, e a saída, em que um resultado é produzido.

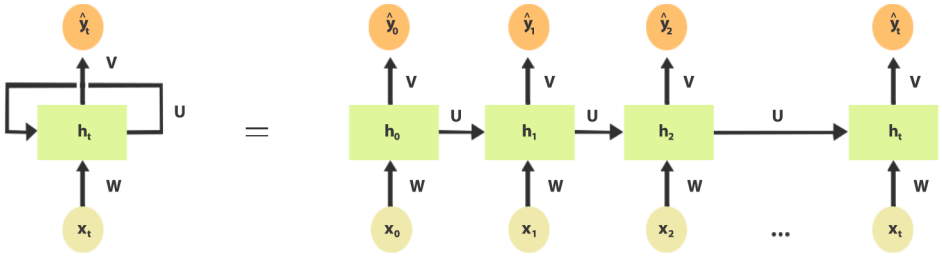
No treinamento das FFNNs, o objetivo está em aprender os parâmetros W e b, o vetor de pesos e vieses, respectivamente, para cada camada, de modo que a saída $\hat{y}$ de cada observação esteja mais próxima possível da verdadeira classe (y). Dessa forma, visando calcular a distância entre $\hat{y}$ e y, é necessário considerar uma função de custo, como a entropia cruzada. Ademais, é aplicado o algoritmo de otimização, como gradiente de descida, para encontrar parâmetros que minimizem essa função de custo.

Devido ao grande número de parâmetros nas várias camadas de sua arquitetura, surge a dificuldade em computar os pesos das camadas iniciais, uma vez que a função de custo está associado às últimas. O algoritmo de *backpropagation* soluciona essa questão com a regra da cadeia, que parte da camada de saída em direção à de entrada, aplicada para encontrar o erro (RUMELHART; HINTON; WILLIAMS, 1985).

2.3.3 Redes neurais recorrentes (RNN)

As redes neurais recorrentes (ou RNN, do inglês *Recurrent Neural Networks*), introduzidas em McClelland *et al.* (1986), são aquelas que contêm conexões cíclicas, i.e., o cômputo das suas unidades depende dos valores de saída dos passos de tempos anteriores. Por conta deste aspecto cíclico, a RNN é bastante aplicada para processamento de dados sequenciais, especialmente para textos (NIELSEN, 2015). A Figura 1 ilustra a arquitetura de uma RNN simples (ELMAN, 1990).

Figura 1 – Arquitetura de uma simples RNN



Fonte: Alex Gwo Jen Lan, 2022

Como observado na Figura 1, a inferência de uma RNN simples é similar aos de FFNNs. Para cada passo de tempo t, é tomado um vetor de entrada x_t para computar y_t, usando o valor de ativação da camada escondida h_t. Nota-se que o valor de h_t também é propagado para o próximo passo $t+1$. Além disso, os parâmetros W, U e V correspondem aos vetores de peso presentes na rede. A seguir, são apresentadas duas equações que descrevem o cálculo de h_t e y_t.

$$h_t = g(Uh_{t-1} + Wx_t)$$

$$\hat{y}_t = f(Vh_t)$$

Como descrito na primeira equação, o estado escondido de cada tempo t (h_t) é calculado com uma função de ativação g aplicada sobre a soma do produto entre a matriz de peso W e a entrada x_t com o produto entre a matriz de peso U e o estado escondido anterior h_{t-1}. Na etapa seguinte, descrita pela segunda equação, a predição $\hat{y}$ é calculada usando uma função de ativação f sobre o produto entre a matriz de peso V e o estado escondido h_t.

As arquiteturas de RNNs são bastante flexíveis e com elas, podem ser projetadas diferentes tipos de variações como as empilhadas e as bidirecionais (SCHUSTER; PALIWAL, 1997). Na primeira variação, múltiplas redes são empilhadas, de forma que o resultado produzido por uma camada, serve como entrada para à sua subsequente. Nas RNNs bidirecionais, a computação flui em ambas direções, de forma independente e simultânea.

A computação das RNNs bidirecionais é similar a de uma recorrente simples. Neste caso, suponha que h_t^f seja o valor produzido no estado escondido no percurso *forward*

e h_t^b, no *back*. Logo, para um passo de tempo t, a predição $\hat{y}$ pode ser definida como a combinação dos estados escondidos desses dois tipos de fluxos, como mostrado na equação adiante. Essa combinação pode ser feita com a concatenação de ambos, e inclusive, com a adição ou multiplicação a nível de cada um dos seus elemento.

$$\hat{y}_t = g(W[h_t^f, h_t^b])$$

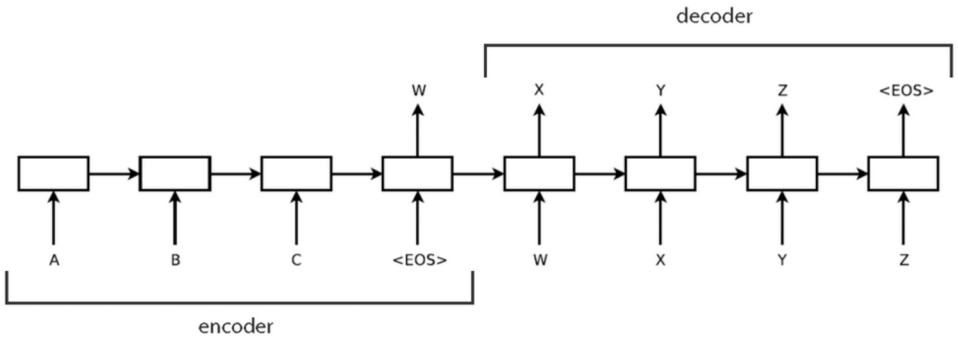
As RNNs apresentam dificuldades em controlar as dependências quando recebem sequências de entradas muito longas (BENGIO; SIMARD; FRASCONI, 1994). Nessas situações, a informação dos passos de tempo iniciais possuem muito menos influência nas saídas da rede. Isso é devido ao fato de que durante o treinamento das RNNs, tendo um gradiente chegado de volta às camadas iniciais como produto de todos os termos a partir das camadas posteriores, é propício em termos uma situação instável, especialmente se os valores se tornam muito pequenos, ocasionando o "desaparecimento do gradiente".

Assim, para solucionar este problema, projetou-se uma variação de RNN conhecida como *Long Short-Term Memory* (LSTM) (HOCHREITER; SCHMIDHUBER, 1997). A sua arquitetura é composta por estruturas de memória e portões que controlam as informações, mantendo o necessário para o contexto. Uma outra alternativa, é o *Gated Recurrent Units* (GRU), que simplifica o treinamento da LSTM, com o uso separado do vetor de contexto e redução do número de portões (CHO *et al.*, 2014).

2.3.4 Redes *encoder-decoder*

As redes *encoder-decoder*, ou modelos *sequence-to-sequence*, são compostas por dois módulos principais. Como observado na Figura 2, o lado esquerdo da rede, o *encoder*, é responsável por gerar uma representação contextualizada a partir de uma sequência de entrada. Em seguida, essa representação é passada para o lado direito da rede, o *decoder*, para a geração de uma sequência de saída de uma tarefa específica (SUTSKEVER; VINYALS; LE, 2014).

Figura 2 – Arquitetura de uma rede *encoder-decoder*



Fonte: Adaptada de Sutskever, Vinyals e Le (2014)

Para a atuação como *encoder*, podem ser consideradas arquiteturas de redes convolucionais, simples RNNs, LSTMs, GRUs e redes de *transformers*. Além do mais, são comumente usadas as Bi-LSTMs empilhadas. Na última camada escondida do *encoder*, é gerado o vetor de contexto w, que refere-se à sequência de representações contextualizadas da entrada (JURAFSKY; MARTIN, 2019).

No *decoder*, o vetor w é recebido em sua primeira camada escondida. O processamento dessa representação resulta em uma sequência de saída, na qual seus elementos são gerados uma por vez a cada passo de tempo. Geralmente, a sua arquitetura é a mesma utilizada no *encoder*. Considerando a escolha de LSTM ou GRU, a geração da sequência de saída pode ser descrita pelas seguintes equações a seguir, sendo que os sobrescritos h^e e h^d se referem respectivamente aos estados escondidos do *encoder* e do *decoder*, g uma função de algum tipo de RNN e $\hat{y}_{t-1}$ como a saída amostrada do *softmax* da etapa anterior.

$$w = h^e$$

$$h_0^d = w$$

$$h_t^d = g(\hat{y}_{t-1}, h_{t-1}^d)$$

$$z_t = f(h_t^d)$$

$$y_t = softmax(z_t)$$

Como pode ser observado, o vetor de contexto w é diretamente disponível ao *decoder* apenas em sua primeira camada escondida. Dessa forma, a influência do vetor w é

diminuída à medida que a sequência de saída é produzida. Para contornar essa situação, pode ser empregado o mecanismo de atenção. Nesta abordagem, considera-se um vetor de tamanho fixo w_i, representando todo o contexto do *encoder*, e que atualiza a cada passo i do *decoder*.

2.4 Métodos de representação textual

A modelagem de documentos de texto pode ser um obstáculo por conta do seu formato não estruturado. Os algoritmos de aprendizado de máquina, por exemplo, não conseguem processar diretamente o texto bruto. Nesse sentido, são necessários métodos capazes de representar estes documentos num formato mais acessível. A seguir, são apresentadas brevemente algumas das abordagens mais populares.

2.4.1 Modelo *Bag-Of-Words* (BoW)

Um dos modelos de representação textual mais populares na área de PLN é o *Bag-Of-Words* (BoW). Neste tipo de modelo, um texto de entrada é representado por um vetor que descreve a presença de cada palavra nesse documento, seja na forma de indicadores binários, ou contagem dos *tokens* ou ponderações de frequência, com tamanho igual ao total de palavras únicas do vocabulário. O BoW é amplamente adotado pela facilidade de implementação, assim como a sua flexibilidade, e por solucionar diversos problemas de classificação de texto (SAGI; DEHGHANI, 2014; DEHGHANI, 2016; KAUR; SASAHARA, 2016; TEERNSTRA *et al.*, 2016).

A abordagem do BoW apresenta algumas desvantagens. Esta representação não considera a ordem em que as palavras ocorrem no documento, e nem as relações sintáticas entre si. Além disso, pode se tornar uma representação esparsa, consequência de situações em que a quantidade de palavras do vocabulário é excessivamente grande quando comparado com o número de documentos analisados.

Uma observação importante sobre o modelo BoW é que as estimativas de frequência não são necessariamente bons indicadores de importância das palavras. Por exemplo, palavras da classe de preposições são muito comuns, mas apresentam pouco ou nenhum significado. Por outro lado, palavras raras como “linguagem” podem ser muito mais

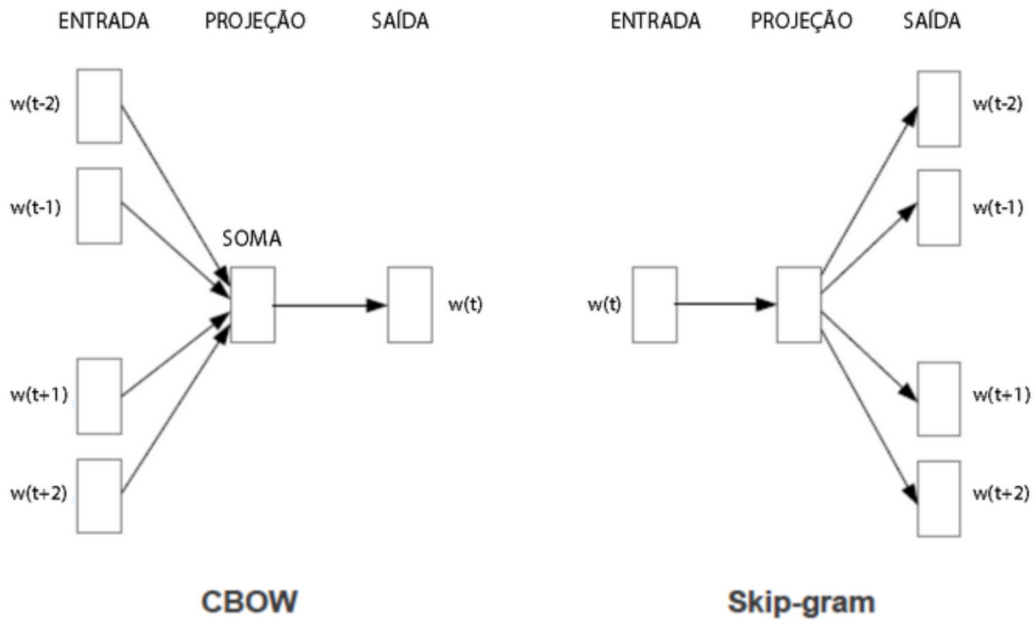
informativas quanto ao significado do texto. Assim, a estatística de TF-IDF é usada para atribuir peso relativo às palavras em função de sua distribuição no conjunto de documentos.

2.4.2 Modelos de representação distribuída de palavra (*word embeddings*)

Como forma de contornar os problemas comumente associados à abordagem de BoW, pode se considerar os modelos de representação distribuída de palavra (*word embeddings*). Este modelo segue a ideia de que o significado de uma palavra é determinado por outras que frequentemente a acompanham, e assim definem o seu contexto. Nesta abordagem, as palavras são representadas em um espaço vetorial de números reais (*word embeddings*) de modo que seja possível posicionar cada uma delas nesse espaço, permitindo a captura de relações sintáticas e semânticas (MIKOLOV; YIH; ZWEIG, 2013).

Para a obtenção dessas representações, destacaram-se os algoritmos *Word2Vec* (MIKOLOV *et al.*, 2013a). Nesse arcabouço para o aprendizado de vetores de palavras são implementadas arquiteturas como a de *Skip-gram* e a de *Continuous Bag-Of-Words* (CBOW). A diferença entre as duas está na forma como os *embeddings* são computados. Na arquitetura CBOW, o objetivo do modelo é predizer uma palavra a partir do seu contexto e, no caso da *Skip-gram*, o contrário. Estes procedimentos são ilustrados na Figura 3.

Figura 3 – Arquitetura do modelo *Word2Vec*



Fonte: Adaptada de Mikolov *et al.* (2013a)

Como pode ser observado na Figura 3, a arquitetura de *Skip-gram* consiste em uma rede neural rasa constituída por uma camada de entrada, uma escondida (ou projeção) e uma de saída. Para o seu funcionamento, considera-se um córpus na qual cada palavra do seu vocabulário fixo é representada por um vetor.

As palavras são inicializadas com vetores de valores aleatórios. É feita uma grande iteração dentro do algoritmo passando-se em cada termo do texto. Localizado uma "palavra central", passa-se pelas palavras ao redor dela (palavras-contexto), e com as predições realizadas, a representação em vetor é alterada com base nelas de maneira a alcançar um melhor resultado.

Durante o treinamento do modelo, o *Skip-gram* visa encontrar representações de palavras mais adequadas para predizer as palavras-contexto. Em outros termos, o objetivo consiste em maximizar a fórmula descrita abaixo (MIKOLOV *et al.*, 2013b), em que T é o tamanho do vocabulário, w é a palavra central e, c é o valor de uma janela de tamanho fixo (por exemplo, consideramos 2 palavras-contexto, um para cada lado $[-2 <= j <= 2]$, totalizando 4 palavras a serem preditas).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<=j<=c, j\neq 0}\log p(w_{t+j} \mid w_t)$$

Observa-se que o termo $1/T$ é incluído à frente da fórmula, para se trabalhar com a média, que auxiliará a manter a escala, e que o logaritmo é aplicado para tornar os produtos na soma de logaritmos dessa probabilidade, evitando o *underflow* numérico. Além disso, a probabilidade p é apresentada com mais detalhes na função a seguir (MIKOLOV *et al.*, 2013b). A fórmula de p contém o produto entre dois vetores w_O e w_I tanto no numerador quanto no denominador, sendo que w_o e w_i correspondem respectivamente à palavras de contexto e central, e W é o número de palavras do vocabulário.

$$p(w_O \mid w_I) = \frac{exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} exp({v'_{w_O}}^T v_{w_I})}$$

O produto do numerador é um indicativo de similaridade entre as palavras w_O e w_I. Nessa multiplicação, valores maiores são alcançados entre números que apresentam o mesmo sinal para componentes das posições correspondentes, e na situação contrária, valores menores, quando operam-se números de sinais opostos e magnitudes diferentes. No numerador, é utilizado uma função exponencial para tornar o resultado positivo, o que

se torna necessário por se tratar de uma probabilidade. Como um todo, trata-se de uma função *softmax*, que mapeia qualquer número em uma distribuição de probabilidade.

A formulação de $p(w_O \mid w_I)$ se mostra impraticável devido ao alto custo de computar o gradiente de $\log p(w_O \mid w_I)$, que seria proporcional à quantidade de palavras no vocabulário (W). Dessa maneira, em Mikolov *et al.* (2013b) é proposta uma possível solução com a técnica de *negative sampling*.

A principal ideia de *negative sampling* é treinar regressões logísticas binárias de modo a atribuir um grande valor de probabilidade para as palavras-alvo verdadeiras (w_t), e, por outro lado, designar baixas probabilidades para as k-palavras ruídos aleatoriamente amostradas, também denominadas "*negative samples*". Assim, a velocidade de treinamento se torna mais rápida, uma vez que a função escala conforme o valor de k.

No caso do método CBOW, é realizado o processo inverso ao do *Skip-gram*. O objetivo da tarefa é predizer uma palavra baseada em outras que se encontram próximos a ela dentro de um limite definido. Entretanto, o processo de treinamento acima também se aplica ao CBOW, e a sua arquitetura também consiste em uma rede neural rasa de três camadas, como mostrado na Figura 3.

Com os vetores de palavras resultantes, é possível avaliá-los intrinsecamente com analogias sintáticas ou semânticas (MIKOLOV; YIH; ZWEIG, 2013). A primeira se refere à avaliações com, por exemplo, tempos verbais, plurais, adjetivos comparativos, como em "verei está para ver assim como falarei está para x", em que x é a representação gerada. Nas analogias semânticas, considera-se relações como "França está para Paris, assim como Itália está para x". As analogias também podem ser expressas como operações entre os vetores, como no popular exemplo vetor(rei) - vetor(homem) + vetor(mulher) = vetor(rainha).

2.4.3 Modelos de *embeddings* dependentes de contexto

Com a abordagem de representação distribuída descrita na subseção anterior, é obtido apenas um tipo de *embedding* por palavra, independentemente do contexto em que ela ocorre. Em algumas aplicações de PLN, essa distinção pode ser essencial, visto que uma mesma palavra pode ter diferentes comportamentos sintáticos, semânticos e gramaticais.

Os *embeddings* dependentes de contexto podem ser produzidos com o uso de modelos de língua pré-treinados como ELMo (*Embedding from Language Models*) (PETERS *et al.*,

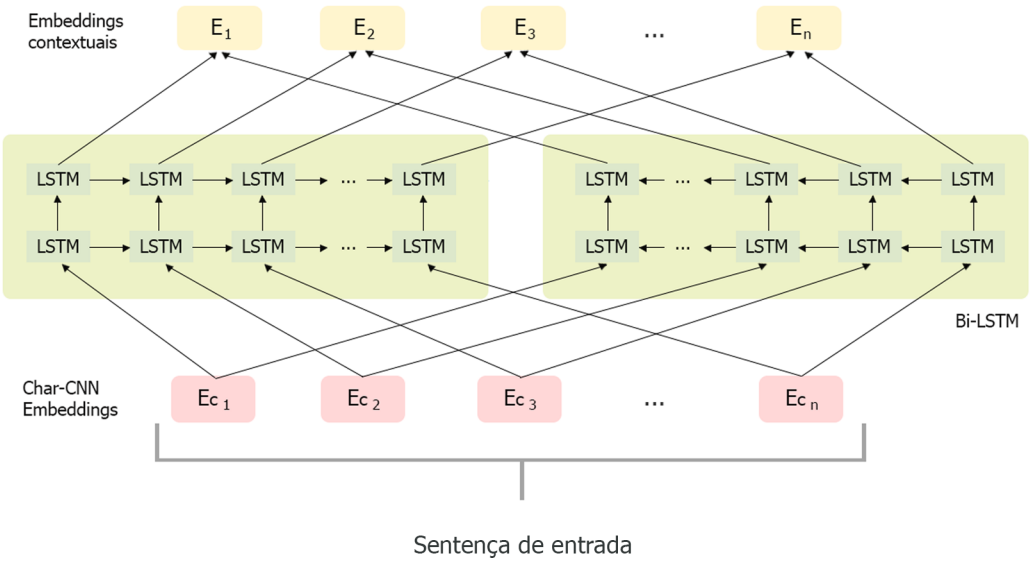
2018) e BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN *et al.*, 2019). Estes modelos têm apresentado resultados superiores em diversas tarefas de classificação, como Joshi *et al.* (2019), Pecar, Simko e Bieliková (2019) e Zampieri *et al.* (2019). Assim, considerando o exemplo da palavra "banco", que pode ser interpretado como instituição financeira ou mobília, a principal ideia é que teríamos um vetor único na representação estática (*Word2Vec*), enquanto que com *embeddings* contextuais, teríamos vetores distintos para cada contexto de uso desta palavra. Estes modelos já apresentam versões treinadas a partir de córpus em português brasileiro (CASTRO, 2019; SOUZA; NOGUEIRA; LOTUFO, 2020) e serão utilizados na presente pesquisa.

ELMo

ELMo (*Embeddings from Language Models*) (PETERS *et al.*, 2018) é um modelo de língua sensível ao contexto que verifica integralmente uma sentença de entrada antes de produzir o *embedding* de uma palavra. Dessa forma, são adquiridas informações de contexto que podem permitir a geração de diferentes representações para uma mesma palavra dependendo do seu comportamento sintático e semântico dentro da sentença.

A arquitetura do modelo é constituída por uma camada de codificação de caracteres, uma camada de convolução desses caracteres e duas camadas de LSTM bidirecionais, como apresentado na Figura 4. Essas últimas camadas, as LSTMs bidirecionais, promovem uma maior quantidade de informações contextuais para a geração da representação, diferente do modelo *Word2Vec* em que o contexto estava limitado a uma janela de tamanho fixo.

Figura 4 – Arquitetura do modelo ELMo



Fonte: Alex Gwo Jen Lan, 2022

Como pode ser observado em sua arquitetura, uma sentença de entrada é passada para uma camada de codificação de caracteres. Essa camada possibilita o modelo ELMo a trabalhar com palavras ausentes no conjunto de dados de treinamento, tornando a aplicação mais robusta. Em seguida, uma camada de convolução é responsável por gerar a partir dos resultados da camada anterior um *embedding* único e de tamanho fixo.

Assim, depois da etapa de convolução, o *embedding* resultante é processado por duas camadas de LSTMs bidirecionais. O processamento de cada uma dessas camadas pode ser interpretado como se o modelo ELMo utilizasse uma LSTM para percorrer da direita para esquerda a sentença de entrada codificada e outro LSTM, da esquerda para direita. Assim, o modelo considera tanto as palavras que antecedem quanto os que procedem o termo que está sendo predito. Essa estratégia é adotada para garantir mais informações de contexto. Segundo os autores, acredita-se que a primeira camada de LSTM captura melhor as propriedades sintáticas, e a segunda, as semânticas.

BERT

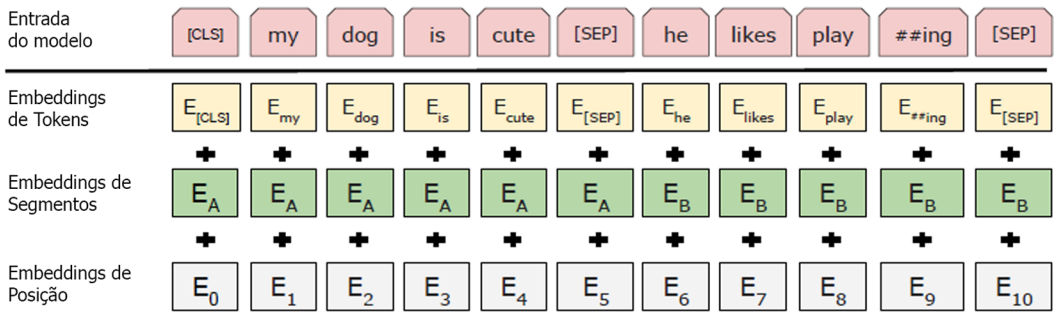
BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN *et al.*, 2019) é uma arquitetura de redes neurais baseada em *Transformers* para a geração de

embeddings contextuais. Diferente dos modelos anteriormente apresentados, o processo de geração desses *embeddings* não é realizado por meio do treinamento de um modelo de língua, i.e., os *embeddings* não são gerados a partir de uma tarefa de predição da próxima palavra com base nos termos que o antecedem.

O modelo BERT é treinado em duas etapas. A primeira, denominada etapa do modelo de língua mascarado, trata do mascaramento aleatório de 15% das palavras de uma sentença, e com isso, o modelo deve predizer estas palavras com base nas outras presentes na sentença. A segunda etapa consiste em predizer se uma dada sentença B sucede uma sentença A.

O modelo recebe um conjunto de sentenças ao qual são extraídas três tipos de informações para cada *token*, como é mostrado na Figura 5. Os *embeddings* de posição são os que permitem indicar onde uma palavra se encontra dentro de uma sentença. Os *embeddings* de segmento mostram se uma determinada palavra pertence a uma sentença A ou B. Os *embeddings* de *tokens* correspondem a uma representação inicial de cada termo, incluindo *tokens* como "CLS" e "SEP" que indicam respectivamente o início e o final de uma sentença. As três informações combinadas, constituem a entrada do modelo.

Figura 5 – Entrada do modelo BERT



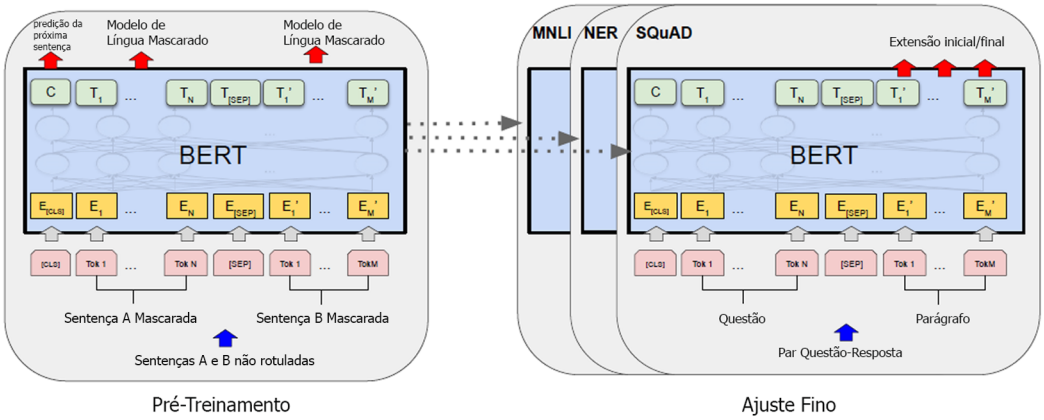
Fonte: Adaptada de Devlin *et al.* (2019)

O processamento do BERT é representado na Figura 6. Como observado no módulo à esquerda (pré-treinamento), o modelo recebe *embeddings* E_i como entrada, que são passadas para blocos de *Transformers*, resultando em outros *embeddings* T_i. Dessa forma, cada representação T_i é utilizada na etapa de modelo de língua mascarado e também na etapa de predição da próxima sentença.

No módulo à direita (ajuste fino) da Figura 6, temos a aplicação dos parâmetros provenientes do pré-treinamento na inicialização de diferentes tarefas de PLN como testes

de perguntas e respostas com *Stanford Question Answering Dataset* (SQuAD), *Name Entity Recognition* (NER) e *Multi-Genre Natural Language Inference* (MNLI). Estes parâmetros são ajustados conforme os dados rotulados de cada tarefa.

Figura 6 – Arquitetura do modelo BERT



Fonte: Adaptada de Devlin *et al.* (2019)

A restrição dos modelos BERT e suas variações está no alto custo computacional de treinamento, ao qual exige-se uma robusta infraestrutura, muitas vezes inacessível devido ao alto investimento financeiro dos recursos computacionais. Por outro lado, para se tornar mais praticável a utilização do BERT, modelos pré-treinados são disponibilizados. Dessa forma, é possível aplicar o processo de transferência de aprendizado.

Com a transferência de aprendizado, torna-se viável o aproveitamento dos parâmetros de um modelo pré-treinado para o desempenho de outras tarefas similares, conforme visto no módulo de ajuste fino da figura anterior. As principais vantagens alcançadas por este processo incluem a diminuição do tempo necessário para treinar o modelo, além da possibilidade de ajustar tarefas em conjunto de dados menores, uma vez que o modelo já aprendeu o suficiente a partir de outras tarefas, e por fim, o ajuste melhora as predições por conta do conhecimento adquirido na etapa de pré-treinamento.

3 Trabalhos relacionados

A revisão bibliográfica foi conduzida de forma exploratória e priorizou trabalhos mais recentes relacionados à classificação de fundamentos morais a partir de texto. Os mesmos foram selecionados a partir de repositórios como a ACL *Anthology*, ACM, IEEE e do *site* oficial da *Moral Foundations*[1]. A revisão é particionada em duas categorias: abordagens baseadas no Dicionário de Fundamentos Morais (GRAHAM; HAIDT; NOSEK, 2009) e abordagens orientadas a dados. Ao final do capítulo, são apresentadas uma seção para as aplicações computacionais e outra para as considerações obtidas com a revisão.

3.1 Abordagens baseadas no Dicionário de Fundamentos Morais

As abordagens baseadas no Dicionário de Fundamentos Morais seguem a ideia de uma análise de conteúdo quantitativo de textos como forma objetiva para tratar dados linguísticos (SMITH, 2006). Nesses trabalhos, é aplicada a técnica introduzida em Graham, Haidt e Nosek (2009) e que consiste na classificação a partir de texto usando o número de correspondências entre as palavras do documento analisado e os termos presentes no DFM.

Em Clifford e Jerit (2013) é realizada uma investigação da retórica moral entre proponentes e opositores da pesquisa de células-tronco. Para isso, são analisados manualmente artigos da *New York Times* numa cobertura de 12 anos (1999-2010). Com esse estudo, concluiu-se que os proponentes têm se concentrado quase exclusivamente no fundamento Cuidado (vício) para reforçar a sua posição, enquanto que os opositores empregaram esse mesmo fundamento e, em menor grau, o de Pureza.

Ao contrário do trabalho apresentado em Clifford e Jerit (2013), num dos estudos apresentados em Graham, Haidt e Nosek (2009) foi utilizado o dicionário LIWC (TAUSCZIK; PENNEBAKER, 2010), ao invés de uma análise manual. Este método objetivou a compreensão da diferença do discurso moral entre liberais e conservadores dos Estados Unidos. Para este fim, consideraram-se sermões das igrejas *Unitarian* (69 sermões) e *Southern Baptist* (34 sermões), entre 1994 e 2006. Como resultado, afirma-se que liberais tendem mais à utilização das palavras ligadas ao Cuidado e Justiça, enquanto conservadores tendem mais a termos associados à Lealdade e Autoridade.

[1] moralfoundations.org

De forma similar, em Sylwester e Purver (2015) são exploradas as diferenças psicológicas, que vão além dos valores morais, entre indivíduos com orientações políticas divergentes. Com o uso do dicionário LIWC, analisou-se *tweets* dos 5.373 seguidores democratas e dos 5.386 seguidores republicanos. Ao final, observou-se que o primeiro grupo tende mais ao uso de 1ª pessoa do singular, o que enfatizaria o seu caráter de "individualidade", e também de palavras que expressam ansiedade e sentimentos. Por outro lado, é sugerido também que seguidores republicanos tendem a termos que valorizam a ideia de grupo, de promessas e de religiosidade.

Sob outra perspectiva, em Fulgoni *et al.* (2016) foram identificados os diferentes termos que liberais e conservadores empregavam para discutir um mesmo tema de diversas fontes de notícias (e.g., mudanças climáticas, aborto, violência policial e etc.). Esse trabalho obteve a frequência das palavras por meio do *Media Cloud API*. Segundo os resultados, que confirmam a conclusão dos trabalhos anteriores, liberais apresentam mais termos de Cuidado e Justiça (vício e virtude), enquanto os conservadores focaram em Autoridade e Lealdade, tratando-se mais da violação dos mesmos (vício).

A técnica baseada em DFM é simples e efetiva, como demonstrado nos trabalhos desta seção. No entanto, por depender de um dicionário, tem-se alguns desafios. Primeiro, existe uma dificuldade em se capturar o significado das palavras em todos os contextos possíveis. Segundo, uma mesma palavra pode ter diferentes conceitos, inclusive opostos (e.g., positivo ou negativo) dependendo do contexto. Terceiro, os termos do dicionário podem surgir e desaparecer, além de mudarem seus significados ao longo do tempo. Para solucionar desafios deste tipo, podem ser usadas abordagens orientadas a dados.

3.2 Abordagens orientadas a dados

Diferentemente das abordagens baseadas em DFM, as que são orientadas a dados tendem a não depender tanto do recurso do dicionário. Assim, trabalhos dessa abordagem classificam fundamentos morais com base em padrões textuais, mediante aplicação de métodos de aprendizado de máquina. Estes métodos são empregados devido à capacidade de processar grandes volumes de dados, possibilitando a análise de postagens de redes sociais no estudo da classificação de fundamentos morais.

3.2.1 Análise Semântica Latente para quantificação de moralidade

É sugerida em Sagi e Dehghani (2014) uma abordagem para mensurar a moralidade em textos dentro de uma configuração mais natural e automática, em comparação ao método tradicional discutido na seção anterior. A quantificação é definida pelo uso da técnica de Análise Semântica Latente (LSA, do ingês *Latent Semantic Analysis*) (DEERWESTER *et al.*, 1990) para representação dos textos analisados.

Como primeiro passo desta técnica, é realizada a tabulação do córpus de interesse em uma matriz de co-ocorrência entre palavras. Cada célula desta estrutura contém a frequência de uma palavra-chave (linha) em relação a um termo (coluna) dentro de uma janela de tamanho 15. Em seguida, é empregada a decomposição de valor singular para converter a matriz num espaço semântico na qual as N dimensões mais importantes são usadas. Para este trabalho, são consideradas 100 dimensões. Dentro do espaço, cada palavra é representada por um vetor.

A partir dos vetores do espaço semântico, é possível gerar a representação do significado de um grupo de palavras, denominado vetor de contexto. Para isso, é realizada a adição de todas as representações vetoriais das palavras que compõem o grupo. Existem dois grupos de vetores de contexto, sendo que um deles é computado a partir das palavras dos documentos analisados, e o outro é calculado a partir de um conjunto de termos moralmente relevantes do córpus aos quais estão associados a um fundamento específico e se encontram dentro do DFM.

Como próximo passo, é medida a similaridade entre os dois grupos de vetores de contexto. Para este fim, pares de vetores são randomicamente amostrados, um de cada grupo e sem repetição para o cálculo da similaridade de cossenos. Esse processo é repetido até o esgotamento dos vetores disponíveis num dos grupos. A seguir, é feita a média dos cossenos para definir uma medida geral da similaridade entre ambos agrupamentos. Esta operação é realizada 1.000 vezes, para fins de estabilidade, tomando a média delas como resultado final. Assim, tem-se um único número que representa o quanto os documentos estão associados a um dos fundamentos morais.

Para testar esse método, são apresentados três estudos de casos. No primeiro deles, com uma coleção de 1.8 milhões de artigos da *New York Times* (entre janeiro de 1987 a junho de 2007), analisou-se a influência dos ataques terroristas ao *World Trade Center* na

carga moral das discussões sobre esta edificação. No segundo, considerou-se um compilado de 3.449 postagens de *blogs* sobre o *Ground-Zero Mosque* (2010) para identificação de cargas morais nos debates entre liberais e conservadores. No terceiro, cerca de 230.000 transcrições dos discursos do Senado dos EUA (1989-2006) foram usados para compreender o posicionamento moral dos senadores democratas e republicanos na questão do aborto.

A partir destes experimentos, foi comprovado que o método LSA pode capturar as dimensões morais desses eventos, contribuindo assim para uma análise que valorize o significado das palavras em detrimento da frequência delas. Além disso, o método pode ser aplicado para diferentes tipos de textos (artigos, *blogs* e discursos). Em vista desses resultados, a mesma abordagem também é empregada em Dehghani (2016), na qual se propõe a identificação de valores morais que atuam como fatores de distância social entre indivíduos.

Neste segundo trabalho, foi calculada a carga moral das palavras extraídas de um córpus de 731.332 *tweets* escritos por 220.251 usuários sobre o evento político *U.S. Federal shutdown*, ocorrido em 2013. Como observado, existe um aumento na diferença da pontuação de Pureza à medida que temos uma maior distância nas conexões, e concluiu-se que tal fundamento é o que mais influencia nessa aproximação quando comparado com os outros quatro.

3.2.2 Análise Semântica Latente para quantificação de moralidade em textos curtos

Com base no algoritmo de LSA, o trabalho apresentado em Kaur e Sasahara (2016) investiga o papel de cada um dos cinco fundamentos morais e do relacionamento entre eles nas discussões cotidianas em plataforma de mídia social. Para isso, os fundamentos presentes nos textos são mensurados utilizando-se *tweets* associados com temas em que há possibilidade de violação moral.

Os conjuntos de dados são córpus extraídos a partir das postagens do Twitter em inglês no período compreendido entre 1º de março a 24 de abril de 2016. Os córpus e seus respectivos tamanhos são aborto (1.516.119 *tweets*), homossexualidade (456.674 *tweets*), imigração (210.286 *tweets*), religião (462.8102 *tweets*), e imoralidade em geral (217.975 *tweets*). Apenas este último é para a análise textual, e os demais são processados para a identificação de palavras-chaves e de contexto associados a estes tópicos.

A quantificação dos fundamentos é realizada com uma adaptação do modelo de LSA utilizada em Sagi e Dehghani (2014) e Dehghani (2016), discutidos na seção anterior. A diferença é que antes da transformação da matriz com a decomposição do valor singular, a mesma passa a ser convertida em uma matriz de *positive pointwise mutual information* (PPMI) a fim de evitar a atribuição de pesos altos à palavras mais frequentes que sejam irrelevantes aos assuntos morais.

Nos resultados, considerando que a classificação de um *tweet* é determinada pelo maior valor obtido na similaridade de cossenos, observou-se que o Cuidado é o fundamento dominante nas discussões de imoralidade (21.135 *tweets*), enquanto que a Autoridade é a menos presente (4.932 *tweets*). Para o relacionamento entre os fundamentos, a medida da similaridade dos cossenos é calculada entre eles. Com base nesses resultados mostrou-se também que há uma forte correlação no emprego de Lealdade e Autoridade, e que, por outro lado, a Pureza é o que não apresenta relação significativa com os demais fundamentos.

3.2.3 Representação *Doc2Vec* para textos curtos

Os estudos em Nokhiz e Li (2017) analisam a influência dos fundamentos morais sobre o comportamento de avaliação dos usuários de mídias sociais, com ênfase em contextos de imoralidade. É proposta uma abordagem com *Doc2Vec* (LE; MIKOLOV, 2014) para representar os dados textuais da plataforma de avaliações *Yelp*.

As avaliações são extraídas a partir de um conjunto de dados aberto do *Yelp Dataset Challenge Round 10*, e são referentes a 4.153.151 comentários em inglês sobre vários tipos de negócios empresariais, com uma nota que varia entre 1 a 5 estrelas. A partir desse conjunto, as avaliações foram filtradas com palavras-chaves "*moral*" e "*ethic*", resultando em 7.039 comentários.

Como método de representação desses dados, foi adotado o *Doc2Vec*. Esse modelo cria uma representação vetorial para um conjunto de palavras tomados coletivamente como uma única instância. Mais especificamente, é trabalhado o modelo de vetores de parágrafo com memória distribuída. Assim, cada comentário é considerado como um documento, e convertido em uma representação vetorial de tamanho 100.

Da mesma maneira como é construído um vetor de parágrafos, considerou-se todas as palavras de vício de um fundamento moral por vez para construir a sua representação. Esse

processo é realizado por meio da soma dos vetores de cada um desses termos. Para análise, foram utilizadas as 149 palavras de vício do DFM. Por fim, para medir a aproximação entre um comentário com um dos fundamentos, é calculada a similaridade de cossenos. Esse valor corresponderia à carga moral de cada um dos cinco fundamentos morais.

Num dos estudos de caso, o propósito está em compreender se indivíduos que possuem maior preocupação moral avaliam de maneira diferente quando comparado com os usuários regulares. Assim, como primeiro passo, identificou-se a frequência de avaliações para cada um dos fundamentos.

Para esta etapa inicial, realizou-se os cálculos das cargas morais, e os comentários foram organizados nos respectivos fundamentos de acordo com os valores de saída. Considerou-se um limiar de 0,2 para o cálculo da similaridade de cossenos. Com isso, foram observados que 1.002 comentários pertencem à categoria de Lealdade, 1.115, à Autoridade, 1.118, à Cuidado, 1.188, à Justiça e 1.359, à Pureza.

Em seguida, com essa distribuição de comentários entre os fundamentos, os mesmos foram organizados pela avaliação de estrelas. Devido ao desbalanceamento do conjunto de dados considerou-se o cálculo da frequência relativa condicional. A mesma é determinada pela razão entre a quantidade de avaliações por fundamento e a quantidade de avaliações no conjunto de dados.

Reunindo dois grupos de usuários, os regulares e os que possuem preocupação moral em suas avaliações, compararam-se as suas distribuições da frequência e da frequência relativa condicional. O resultado dessa análise indicou que, para os mesmos negócios avaliados, usuários com preocupação moral tendem a dar mais notas baixas (i.e., uma estrela) do que os regulares, especialmente em casos de violação dos fundamentos.

3.2.4 Dicionário de Representação Distribuída para expansão do DFM

A técnica baseada na contagem de palavras com DFM, descrita na seção 2.1.2, apresenta alguns desafios de aplicação. Dentre estes desafios, destaca-se a questão do uso de textos com tamanhos reduzidos (e.g., postagens de redes sociais). Assim, para viabilizar uma gama mais ampla de estudos sobre estes dados, em Garten *et al.* (2018), é proposto o Dicionário de Representação Distribuída (DRD). Este método explora as relações de

similaridade semântica entre as palavras do vocabulário de um conjunto de dados e os "conceitos-chave" gerados a partir de um dicionário.

No método de DRD, consideram-se como entrada um dicionário não-vazio (e.g., DFM) e um vetor n-dimensional de valores reais pré-treinados, que corresponde à representação distribuída dos termos de um vocabulário (e.g., *Google News* e *Wikipedia*, com *Word2Vec*). A partir dessas entradas, é realizada a intersecção para identificar-se as palavras do dicionário que se encontram no vocabulário. Todas as representações distribuídas resultantes dessa operação são então somadas e normalizadas para se criar os conceitos-chave. Criados os conceitos, é calculada a similaridade de cossenos entre cada um deles com os documentos analisados. Esta fórmula pode retornar valores entre -1 e 1, que indicam, nessa ordem, os graus mínimo e máximo de similaridade entre os termos.

No contexto de estudo da TFM, DRD é avaliado num experimento com uma amostra de 3.000 *tweets* relacionados ao Furacão *Sandy*. O classificador utilizado é uma regressão logística com validação cruzada de 10 partições. Consideraram-se três métodos automáticos para a identificação da retórica moral.

No primeiro método, é usada a tradicional abordagem baseada em DFM. De forma sumária, em cada *tweet* analisado, contabilizou-se o número de termos para todas as 10 categorias morais do DFM (virtude e vício). Com esta abordagem, um *tweet* é representado por um vetor de dimensão 10 cujas posições contêm as respectivas frequências dos termos de cada categoria moral.

O segundo método considerou a aplicação de DRD com representação de conceitos-chave a partir de modelos *Word2Vec* (MIKOLOV *et al.*, 2013a) e *GloVe* (PENNINGTON; SOCHER; MANNING, 2014). Em ambos modelos, conceitos-chave foram gerados para cada uma das 10 categorias presentes no DFM. Para cada *tweet*, é calculada a distância entre ele e as 10 representações de conceito, cujos valores de similaridade são organizados num vetor a ser submetido como entrada ao classificador.

De maneira semelhante à configuração anterior, no terceiro método, usou-se também DRD com a representação de conceitos-chave, porém com uma versão simplificada do dicionário. Para esta versão de DFM, foram selecionados subconjuntos representativos das categorias do dicionário (quatro palavras por categoria), com a consulta dos autores desse recurso léxico.

Ao final, é mostrado que os resultados do segundo e do terceiro métodos, cujos valores são próximos entre si, são significativamente superiores em comparação com o

do primeiro. Porém, é importante ressaltar que o DRD não substitui o método baseado no DFM, mas se mostra eficiente em diferentes tipos de tarefas, como as relacionadas à análise de textos com tamanhos restritos, assim como para o uso de pequenos dicionários.

3.2.5 *Naive Bayes* multinomial e BoW para predição de fundamentos morais

Em Teernstra *et al.* (2016), é apresentada a abordagem *MoralityMachine* para a detecção e o rastreamento de fundamentos morais, em inglês, na plataforma *Twitter*. O intuito deste estudo é determinar se as técnicas de aprendizado de máquina são capazes de classificar *tweets* em fundamentos morais, sem a dependência de dicionários, com uma acurácia satisfatória.

Os dados considerados se referem às discussões da opinião pública sobre medidas de austeridade na Zona do Euro, mais especificamente sobre a saída da Grécia como integrante do grupo, o evento "*Grexit*". Estes dados foram extraídos da plataforma *Twitter*, a partir de uma filtragem com as palavras-chaves "*Euro*" e "*Greece*" durante três períodos em 2015 para acompanhar o desenvolvimento do evento. Ao final, o conjunto de dados totalizou 18.986 *tweets*.

Após a coleta, os dados foram pré-processados com a etapa de conversão das letras em minúsculas, remoção de *emojis*, substituição de *links* de *websites* por "URLs"e inclusão de termos mais frequentes à lista de *stopwords*. Depois disso, um conjunto de 2.000 *tweets*, aleatoriamente selecionados, foi rotulado manualmente com o fundamento moral correto, usando como guia o DFM.

Em seguida, selecionando 100 unigramas e 100 bigramas mais comuns, foram criados três conjuntos de dados. O primeiro deles é apenas o conjunto pré-processado, o segundo está organizado em bigramas, e o terceiro contém as palavras selecionadas a partir do recurso DFM. Para cada um dos três conjuntos, existe uma segunda versão com *stopwords* removidas. Esses dados foram representados com o modelo *Bag-Of-Words*.

A tarefa de classificação foi executada com *Naive Bayes* multinomial e também com regressão logística. Dada as diferentes configurações dos conjuntos de dados, foi realizada uma série de experimentos, dentre os quais destacou-se o modelo de *Naive Bayes* sobre dados brutos, com uso de unigrama, e sem atributos do DFM, utilizando validação cruzada de 5 partições.

Por fim, o modelo foi treinado a partir das 2.000 instâncias rotuladas e usado para classificar os outros 16.986 *tweets*. Para cada um dos três períodos do evento, houve o predomínio de um dos fundamentos no debate da "*Grexit*": Cuidado para o primeiro período, Autoridade, para o segundo, e Lealdade, para o terceiro.

Um fato interessante apontado por este estudo foi o de que o modelo de melhor desempenho não fez uso de DFM, sugerindo-se inclusive a descontinuidade do recurso pela sua dificuldade de construção e manutenção. Com esses experimentos, os autores acreditam que o modelo de *Naive Bayes* pode ser um ponto de partida para a tarefa de classificação moral em *tweets* com abordagens de aprendizado de máquina.

3.2.6 *Predição de fundamentos morais com base de conhecimento externa*

A proposta em Lin *et al.* (2018) sugere uma abordagem que possibilita a identificação de valores morais implícitos em textos com restrição de tamanho. Nesse estudo, visa-se a melhoria dos resultados de predição dos valores morais com o enriquecimento contextual dos documentos de entrada. O método aplicado consiste na utilização de uma base de artigos da *Wikipedia*.

Numa visão holística, a abordagem adotada no estudo considera que para cada documento de texto analisado (*tweet*) é retornado um vetor que indica o nível de sua associação a cada um dos fundamentos morais. Esse processo é realizado por dois módulos computacionais. No módulo de extração de características textuais, os atributos são selecionados de um *tweet* e codificados em vetores de representação de palavras. Já no módulo de extração da base de conhecimento, é aplicada a técnica de associação de entidades, que trata da etapa de extração de todos os nomes (e.g., pessoas, lugares, associações, empresas e etc.) para a geração de um conteúdo complementar (informações como a sua definição, ofício, partido, propósito, por exemplo) a partir da conexão com uma base externa.

Em ambos módulos, o modelo usado para gerar esses vetores foi o de *Word2Vec*. Ao final tem-se dois vetores, cada um gerado a partir de seu respectivo módulo, e que são, em seguida, concatenados. O vetor resultante é então utilizado junto com um terceiro (que representa o fundamento moral alvo), como entrada de cinco classificadores binários (um pertencente a cada fundamento), de verdadeiro e falso. Um *tweet* será considerado

"não-moral", caso todos seus valores sejam "falsos" para todos os fundamentos. O processo foi abordado com diferentes modelos de aprendizado supervisionado das quais destacou-se a rede LSTM com uma camada totalmente conectada e uma outra acima com *softmax*.

A abordagem foi avaliada com experimentos a partir de um córpus de 4.191 *tweets* aleatoriamente amostrados, contendo *hashtags* relevantes sobre o Furacão *Sandy*. Consideraram-se 3 configurações de experimento: i) apenas a representação de palavras, ii) combinação do item anterior com a base de conhecimento e iii) combinação dos itens anteriores e o uso do DFM.

Como resultado, foi demonstrado que a integração da base de conhecimento melhora significativamente a detecção de valores morais, e que o uso do dicionário não impactou substancialmente no desempenho final. Ainda, realizou-se outro teste com um anotador humano, e como resultado, tanto o modelo quando este anotador apresentaram um desempenho similar na tarefa de identificação de fundamentos morais.

3.2.7 *Probabilistic Soft Logic* para predição de fundamentos morais

Em Johnson e Goldwasser (2018), é proposta uma abordagem para a identificação de moralidade subjacente em discursos políticos nas plataformas de mídias sociais como o *Twitter*. Dentre as contribuições deste trabalho, incluem-se um conjunto de dados anotados para fundamentos morais, uma orientação sobre este processo e modelos *Probabilistic Soft Logic* (PSL) para a predição desses fundamentos.

Como conjunto de dados, foi utilizado o *Congressional Tweets Dataset* (JOHNSON; JIN; GOLDWASSER, 2017). As postagens foram rotuladas por dois humanos e, para evitar viés político na anotação, cada fator moral é considerado sob a perspectiva do partido político a que pertence o autor do *tweet*. Ademais, é possível atribuir mais de um fundamento para um mesmo *tweet*.

Durante a anotação, observou-se a preferência dos anotadores em atribuir fundamentos com base em pequenas frases ou no *tweet* inteiro, ao invés de ser feito com apenas um termo do DFM. Assim, criou-se uma nova lista de unigramas, denominada "Razão do Anotador" (RA), que é o resultado da compilação de todas as frases dos anotadores por fundamento em um único conjunto.

Além da lista de unigramas RA, um outro tipo de entrada para os modelos PSL são as frases abstratas, que são expressões de alto nível frequentes em discursos políticos. Neste trabalho, considerou-se as seguintes abstrações: legislação ou voto, direitos de igualdade, emoção, fontes de perigo e dano, benefícios e efeitos positivos, solidariedade, ações políticas, proteção e prevenção, valores ou tradições americanas, religião e promoção. Por exemplo, em "O Presidente Obama assina uma conta", existem dois constituintes: "O Presidente Obama assina" e "assina uma conta", correspondendo às abstrações de ações políticas e legislação.

Algumas dessas abstrações estão correlacionadas com os fundamentos morais (e.g., religião está correlacionada com Pureza). Para associar as frases nos *tweets* a essas abstrações, foi utilizado um modelo de representação de similaridade de frases, treinado sobre o "*Paraphrase Database*" (PPDB) (GANITKEVITCH; DURME; CALLISON-BURCH, 2013) e incorporado a uma rede neural convolucional para capturar estruturas de sentenças. Dessa forma, são geradas as representações das abstrações e as pontuações de similaridade de cossenos entre elas e os *tweets*.

Os *tweets* são usados como entrada de modelos de extração de características, produzindo saídas que são adaptadas para a notação de predicados nos modelos de *Probabilistic Soft Logic* (PSL). Estes predicados (e.g., partido, *frame* e *issue*) são combinados em regras probabilísticas de cada modelo, que são construídas incrementalmente sobre as regras do modelo anterior. Na totalidade, foram desenvolvidos treze modelos de PSL. Para obter a predição final, as regras de alto nível sobre as representações relacionais dessas características são passadas para um *hinge-loss Markov random field* (HLMRF).

Com essas configurações, foram realizados experimentos supervisionados com dois tipos de unigramas, pertinentes ao DFM e ao RA. Para cada um desses dois recursos, empregou-se os seguintes modelos: o voto majoritário, e os modelos de PSL propostos, além do *Support Vector Machines* com *Bag-Of-Words* e PSL com *Bag-Of-Words*. Nestes dois últimos, foi utilizado apenas o DFM.

De acordo com os resultados, observou-se que à medida que as informações são acrescentadas em cada modelo de PSL, maior a medida F1 é alcançada, para ambos grupos de unigramas. Assim, dentre os treze modelos de PSL propostos, aquele que considerou todas as características apresentou os maiores valores de medida F1.

3.2.8 *Probabilistic Soft Logic* com contexto para predição de fundamentos morais

Como extensão do trabalho apresentado em Johnson e Goldwasser (2018) da seção anterior, o estudo em Johnson e Goldwasser (2019) demonstrou a serventia da incorporação de informações sociais (i.e., identificação de postagens compartilhadas, redes de contatos e período de publicação das postagens) e comportamentais para o modelo de *Probabilistic Soft Logic* (PSL), empregando-o em configurações supervisionada e não-supervisionada. Com estes novos modelos, buscou-se um esclarecimento das tendências nos discursos políticos ao longo do tempo e do relacionamento com os eventos mundiais.

O PSL descrito na subseção anterior, que é baseado nas características de linguagem, passa a ser o alicerce para a construção de outros três modelos. O primeiro deles é o RETWEETS, que inclui informações de *retweets*, auxiliando na detecção do efeitos da linguagem e as conexões sociais. O segundo é o FOLLOWING, ao qual incrementa com a rede de contatos, permitindo a visualização de relacionamentos dos políticos e os semelhantes padrões de uso dos fundamentos morais. E o terceiro, TEMPORAL, contribui com a verificação de tweets que ocorrem dentro da mesma janela de tempo de um dia.

Estes três novos modelos de PSL são submetidos a experimentos de duas configurações: supervisionada e não-supervisionada. Em ambos, o conjunto de dados empregado é o *Congressional Tweets Dataset* (JOHNSON; JIN; GOLDWASSER, 2017). Além disso, para as duas configurações, o PSL baseado na linguagem é aplicado como *baseline*, visando salientar as melhorias oferecidas pelas alterações adicionais.

Para os experimentos supervisionados, demonstrou-se uma melhoria na predição geral para todos os fundamentos com a inclusão dessas características sociais e comportamentais. Os experimentos foram conduzidos com a validação cruzada com 5 partições. O aumento corresponde a 9,14 pontos na média de valores da medida F1. Com ressalva, informações de *retweet* não contribuíram de maneira significativa, provavelmente, devido à sua escassez nos dados coletados.

Nos experimentos não-supervisionados, na qual os *tweets* são classificados com uma implementação de PSL com algoritmo de expectativa-maximização, observou-se também que as informações sociais e comportamentais elevaram os resultados da predição. Mais especificamente, o modelo combinado final apresentou uma melhoria da média de valor da medida F1 de 12,06 pontos sobre o do *baseline*.

Com os modelos de configuração não-supervisionada, geraram-se predições a partir de duas coletâneas de *tweets* do Senado, o *Senate Tweets 2016* e *CongressTweets*, de 2018. Estes resultados possibilitaram uma análise qualitativa do relacionamento entre os fundamentos morais implícitos no discurso em mídia social e os eventos políticos mundiais. Assim, elaboraram-se dois estudos de caso para investigar melhor esta relação.

Em um desses estudos de caso, com ênfase nas tendências do ano, consideraram-se os seguintes assuntos: direitos da mulher e do LGBTQ, violência armada, imigração, terrorismo, e saúde pública. A partir dos resultados, percebeu-se uma maior predileção do partido Republicano pelo fundamento Cuidado em seus discursos, enquanto que os Democratas optaram por Cuidado e Justiça. Numa perspectiva cronológica, é notado que o fundamento Cuidado é menos utilizado pelos Republicanos, enquanto que os Democratas passam a utilizá-lo, junto com os outros, à medida que mais políticos deste partido começam a discutir no *Twitter*. Para ambos, existe um pico no emprego do fundamento Cuidado nos dias que antecedem eventos eleitorais.

Com este trabalho, comprovou-se a utilidade de informações comportamentais para melhorar o desempenho dos modelos preditivos de PSL em ambas configurações supervisionada e não-supervisionada. Estes modelos podem colaborar para mostrar tendências em discursos políticos ao longo do tempo, assim como a relação deles com os eventos globais.

3.2.9 *MoralStrength* para predição de fundamentos morais

Em Araque, Gatti e Kalimeri (2020) , visou-se o desenvolvimento de uma expansão do DFM original baseado em "*synsets*"do *WordNet*, o *MoralStrength*. É importante mencionar que nesse recurso léxico cada um de seus lemas possui uma pontuação de força ou "valência moral"associada a um determinado fundamento. Embora o objetivo deste trabalho seja diferente do presente estudo, é interessante considerar os modelos utilizados para avaliar o potencial de predição do *MoralStrength*.

O conjunto de dados usados para avaliar os modelos de classificação é o MFTC. O córpus original contém 35.108 *tweets* anotados, entretanto, segundo o estudo, recuperou-se 82% do conjunto original (24.802 *tweets*). Nessa versão do MFTC, temos os seguintes sete tópicos e suas respectivas quantidades de *tweets*: Furacão Sandy (3.478), Protesto de Baltimore (4.174), "*All Lives Matter*"(3.486), "Black Lives Matter"(4.340), Eleições

Presidenciais de 2016 (4.366) e discursos de ódio (4.958). O conjunto de dados MFTC também é aplicado no presente estudo, com a diferença de que foi utilizada a versão original (seção 2.2.1).

O texto foi pré-processado com a conversão de URLs, nomes de usuários e *hashtags* em *tokens* especiais (i.e., "<url>", "<*username*>"e "<*hashtags*>"), e também com a retirada de pontuações, símbolos e números. Em seguida, para avaliar o potencial do *MoralStrength*, foram empregadas três abordagens de extração de características: *Moral Freq*, *Moral Stats* e *SIMON*.

Moral Freq considera a contagem da quantidade de palavras que expressam um fundamento moral específico de forma binária (ausência ou presença). Para definir se uma determinada palavra expressa um dado fundamento, basta verificar se sua pontuação de "valência moral"no *MoralStrength* encontra-se acima de um limiar (mais especificamente, os autores definiram como 5 por conta de propriedades de generalização do léxico moral). A saída deste modelo para cada documento de texto é um vetor de dimensão 10 que contém as frequências para cada extremidade moral (e.g., Cuidado virtude e Cuidado vício).

Moral Stats recebe um texto como entrada e a partir dele é gerado um resumo estatístico da distribuição de valência moral das palavras presentes. Nesse resumo, estão contidas informações como a média, o desvio padrão, a mediana e o valor máximo. No final, este modelo produz um vetor de dimensão 20 constituído de valores estatísticos obtidos com as anotações do léxico.

SIMON (*SIMilarity-based sentiment projectiON*) usa um modelo de *word embedding* pré-treinado para processar a similaridade de cossenos entre palavras do texto analisado e a seleção de palavras do fundamento moral correspondente no *MoralStrength*. Como saída do modelo, temos um vetor de representação que codifica a similaridade de um documento em relação a um fundamento moral.

Para fins de comparação, foram utilizados o modelo de contagem com DFM original (na seção 3.1) e o de unigrama (BoW) como *baselines*. Além disso, consideraram-se métodos formados pela combinação dos modelos *SIMON* (*SIMON* + *Moral Freq*, *SIMON* + *Moral Stats* e *SIMON* + *Moral Freq* + *Moral Stats*), e também o unigrama com as outras abordagens propostas (unigrama + *Moral Freq*, unigrama + *Moral Stats*, unigrama + *SIMON*, unigrama + *Moral Freq* +*Moral Stats*, e assim por diante). Em todos eles, aplicou-se a regressão logística junto com a validação cruzada de 10 partições.

A partir dos experimentos, observou-se que modelos com *MoralStrength* apresentaram resultados em medida F1 superiores aos dos *baselines* e também comparáveis com os métodos de classificação do estado-da-arte. O modelo que teve os melhores resultados em geral foi a combinação unigrama + *Moral Freq*. Dessa forma, comprovou-se a eficácia da utilização do *MoralStrength* para as tarefas de classificação de fundamentos morais.

3.2.10 DRaiL para predição de fundamentos morais em textos

Em Roy e Goldwasser (2021), buscou-se compreender a relação entre o uso de fundamentos morais pelos políticos em redes sociais e o posicionamento deles em relação às questões como controle de armas e imigração. Para realizar as análises propostas neste estudo, é aplicada uma abordagem denominada *Deep Relational Learning* (DRaiL) (PACHECO; GOLDWASSER, 2021) para a captura de moralidade em texto.

O DRaiL é um arcabouço declarativo para predição com aprendizado estruturado. Nesta ferramenta, consideram-se três regras bases $r1$, $r2$ e $r3$, além de uma restrição c para modelar as características e dependências da tarefa, como o texto dos *tweets*, a afiliação política dos autores (Democrata ou Republicano), o tópico (controle de armas ou imigração) e o período de publicação da postagem.

As regras do DRaiL associam as características a um dos fundamentos morais. Na regra $r1$, é interpretado que "Um *tweet* t, possui um rótulo moral m", na regra $r2$, "Um *tweet* t com autores de afiliação i, possui rótulo moral m", e na regra $r3$, "Um *tweet* t com o tópico k, possui rótulo moral m". A restrição c é traduzida como "Se dois tweets possuem o mesmo tópico, são de autores de mesma afiliação política e foram publicados em períodos muito próximos, então, eles possuem o mesmo fundamento moral".

Para cada uma das regras, é associada uma arquitetura de rede neural. As regras $r1$, $r2$ e $r3$, consideraram um modelo BERT para gerar as representações dos *tweets*. Nas regras r2 e r3, as informações de afiliação política e tópico foram representadas com codificação *one-hot*, por meio de uma FFNN, e concatenadas com o texto de *tweets*, cujos *embeddings* já haviam sido representados pelo BERT. Por fim, as representações finais de todas as regras são computadas por uma função de *hinge-loss Markov random field* (HLMRF) para gerar a predição final do texto.

Como conjunto de dados, considerou-se o *Congressional Tweets Dataset* (JOHNSON; JIN; GOLDWASSER, 2017) para os experimentos. Para a análise dos discursos políticos, realizou-se a coleta de uma série de *tweets* não rotulados, escritos por membros do congresso estadunidense no período de fevereiro de 2017 a 2021. Os *tweets* estão relacionados aos tópicos de controle de armas e imigração, e correspondem, respectivamente, a 74.000 e 87.000 instâncias, subdivididos entre os partidos Democratas e Republicanos.

Para avaliação dos modelos baseados em DRaiL, conduziram-se experimentos para a tarefa de classificação de texto em 11 classes (i.e., além dos cinco fundamentos categorizados em virtudes e vícios, considerou-se a classe "Não-moral"). Para isso, usaram-se 2.050 *tweets* anotados do *Congressional Tweets Dataset*, com aplicação de validação cruzada em 5 partições. A avaliação contou com dois *baselines*. O primeiro deles é a tradicional abordagem baseada na correspondência de termos do DFM com as palavras do *tweet* analisado, e o segundo, uma LSTM bidirecional com representação de palavras em *GloVe*.

Ao final, o modelo de DRaiL combinando todas as regras bases $r1$, $r2$, $r3$ e a restrição c obteve o melhor desempenho. Assim, este modelo foi escolhido para ser treinado sobre 10% de *tweets* aleatoriamente selecionados do conjunto de dados do experimento, e usados para predizer os *tweets* não rotulados sobre controle de armas e imigração. Segundo os autores, o estudo não pretendeu avançar o estado-da-arte na tarefa de classificar texto em fundamentos morais, mas buscou uma abordagem para tratar a tarefa com a disposição de um córpus com grande volume de texto não rotulado junto com outras dependências.

A partir das predições destes *tweets* não rotulados, realizou-se a análise da relação entre os fundamentos e os discursos dos políticos. Nos textos sobre controle de armas, identificou-se que Republicanos usam mais a virtude de Lealdade para abordar o assunto, e os Democratas, o vício de Justiça em seus discursos. No assunto da imigração, Democratas usam mais o vício de Justiça, e os Republicanos, as virtudes de Cuidado e Autoridade. Com isso, demonstrou-se que ambos fenômenos podem ser explicados com a TFM.

3.3 Aplicações computacionais

No âmbito computacional, a TFM atua como uma base para modelos de PLN com ênfase nos estudos sobre moralidade. Em especial, a tarefa de predição de fundamentos morais tem contribuído em várias aplicações para a compreensão de diversos fenômenos

sociais e políticos. Diferente dos estudos discorridos anteriormente, esta seção apresenta trabalhos que não possuem como foco a classificação de fundamentos propriamente dita, mas que a emprega como subetapa de seus principais procedimentos ou que estejam relacionadas a esta tarefa.

Em Takikawa e Sakamoto (2017), é apresentado um estudo comparativo dos discursos do Congresso dos EUA e da Assembleia do Japão. Para isso, empregou-se o método da contagem com representação BoW. Nesse estudo, no contexto político estadunidense, é indicado que Republicanos favorecem mais o fundamento de Justiça em seus discursos e menos o de Lealdade, em detrimento aos Democratas. No cenário político japonês, é observado que membros do Partido Democrata Liberal apresentam mais moralidade negativa nos cinco fundamentos em seus discursos quando comparado com os do Partido Democrata Social.

Em Volkova *et al.* (2017), foram desenvolvidos modelos preditivos para classificar postagens de notícias entre as categorias de sátira, fraude, *clickbait* (conteúdo sensacionalista) e propaganda. Segundo os autores, as notícias suspeitas e verificadas apelam para diferentes fundamentos morais de seus leitores. Os resultados indicam que é mais comum o uso de termos morais nos *clickbaits*. Além disso, sugere-se que a sátira emprega mais termos relacionados à virtude do que ao vício do fundamento de Lealdade.

Em Mooijman *et al.* (2018), é sugerida uma associação entre a moralização e os protestos violentos. Por uma série de experimentos comportamentais, é demonstrado que a taxa de retórica moral cresceu nas redes sociais (*Twitter*) nos dias em que ocorreram os protestos em Baltimore. Além disso, também foi observado que a frequência por hora de *tweets* moralmente relevantes predizeram as futuras taxas de prisões durante os protestos.

Em Dinkov, Koychev e Nakov (2019), é proposto um detector de toxicidade em notícias de língua búlgara que desempenha uma classificação multiclasse para notícias falsas, sensacionalistas, discurso de ódio e conspirações, anti-democrático, pró-autoritarismo e difamação. A TFM é utilizada neste trabalho como um dos fatores que distingue tais categorias. Ao final, é mostrado que o modelo obteve valores de acurácia superiores aos modelos *baselines* da tarefa.

Em Wang *et al.* (2019), buscou-se compreender estratégias de persuasão eficazes para convencer indivíduos a realizarem atos de bem social como doação. O objetivo desse estudo é o de estabelecer uma base para o desenvolvimento de sistemas de diálogos persuasivos e personalizados. A partir da coleta de 1.017 diálogos persuasivos de voluntários

de um experimento elaborado pelos autores, foi construído um *baseline* para predizer as 10 estratégias presentes no córpus. Com isso, foi revelado que indivíduos que possuem uma maior inclinação para o Cuidado são mais propícios a realizarem doação.

Em Xie *et al.* (2019), é introduzida uma metodologia para a inferência automática, e em larga escala, das variações dos conceitos das palavras em relação ao seu fundamento moral ao longo de um grande período de tempo. O arcabouço proposto consiste de três camadas: a relevância moral, a polaridade moral e as dimensões morais de granularidade fina. Demonstrou-se a capacidade do arcabouço em inferir essas trajetórias morais a partir de sentimentos em cada um desses três níveis ao longo de um grande período de tempo.

Em Hulpus *et al.* (2020), buscou-se uma abordagem de projeção do DFM em três grafos de conhecimento baseado em *DBpedia*, *WordNet* e *ConceptNet*. Esta projeção resolveria problemas de ambiguidade e da limitação de termos no uso do dicionário. O propósito deste estudo é de pontuar todas entidades e conceitos contidos nesses grafos quanto à sua relevância para cada fundamento moral. A pontuação e a avaliação dessas estruturas de dados foi realizada com o *Personalized PageRank* (PPR) (PAGE *et al.*, 1999). O PPR é uma ferramenta para medir a proximidade de nós em grafos amplamente utilizada em mineração e análise de redes. Dentre os três grafos de conhecimento, o *ConceptNet* se destacou com os maiores valores de acurácia. Como trabalho futuro, os autores buscarão desenvolver ainda mais a abordagem dos grafos para melhorar as predições de fundamentos morais a partir de texto.

Finalmente, em Robertson, Lagus e Kajava (2021), é elaborado um painel que oferece uma visão unificada do sentimento geral sobre a pandemia de COVID-19 reportada pelas notícias de diferentes regiões da Europa, no período compreendido entre janeiro de 2020 e 2021. O conteúdo deste painel considera análises geradas pela classificação das manchetes em sentimento "positivo", "neutro"ou "negativo", e em fundamentos morais. Para este último, as manchetes foram representadas em *document embeddings* e são classificadas para um determinado fundamento de acordo com a proximidade da representação vetorial com relação aos dos termos da DFM. Segundo os autores, esse tipo de ferramenta pode ser integrado em um sistema usado por agências para rastrear tendências de notícias. Além do COVID-19, ele pode ser usado para planejar a cobertura de outros eventos nacionais ou globais como eleições, encontros internacionais e esportivos.

3.4 Considerações

Este capítulo apresentou um resumo de trabalhos recentes sobre classificação de fundamentos morais a partir de texto, e que são mais diretamente relacionados à presente pesquisa. Os trabalhos foram organizados em duas categorias: abordagem baseadas em DFM e abordagens orientadas a dados, e acompanhados de um terceiro grupo de estudos aplicados nesta mesma área.

Nas abordagens baseadas em DFM, observa-se a preferência em se utilizar textos longos, e em inglês, como artigos de jornal e de *blogs*, sermões de igreja, discursos políticos, assim como há experimentações com postagens de rede social. Nota-se também que existe um interesse considerável da comunidade pela aplicação da TFM no estudo de discursos políticos de liberais e conservadores nos Estados Unidos.

Nas abordagens orientadas a dados, os desafios oferecidos pela dependência no DFM são contornados com modelos de aprendizado de máquina. Dessa forma, tornou-se possível análises mais consistentes sobre textos curtos, com preferência pelas redes sociais, especificamente o *Twitter*. Esse fenômeno pode ser explicado pela facilidade de acesso a um grande volume desses dados e também pela popularidade das redes sociais em engajar indivíduos para discussão sobre uma diversidade de assuntos. Os trabalhos da abordagem orientada a dados são sumarizados na Tabela 3 e discutidos a seguir.

Tabela 3 – Resumo dos trabalhos correlatos

Estudo	Idioma	Gênero	Representação	Método
Sagi e Dehghani (2014)	IN	notícias, *blogs* e discursos	LSA	similaridade de cossenos
Dehghani (2016)	IN	*Twitter*	LSA	similaridade de cossenos
Kaur e Sasahara (2016)	IN	*Twitter*	LSA	similaridade de cossenos
Teernstra *et al.* (2016)	IN	*Twitter*	BoW	NB
Nokhiz e Li (2017)	IN	*Yelp*	Doc2Vec	similaridade de cossenos
Garten *et al.* (2018)	IN	*Twitter*	Word2Vec GloVe	regressão logística
Lin *et al.* (2018)	IN	*Twitter*	Word2Vec	LSTM
Johnson e Goldwasser (2018)	IN	*Twitter*	PSL	HLMRF
Johnson e Goldwasser (2019)	IN	*Twitter*	PSL	HLMRF
Araque, Gatti e Kalimeri (2020)	IN	*Twitter*	BoW	regressão logística
Roy e Goldwasser (2021)	IN	*Twitter*	DraiL+BERT +"*one hot*"	HLMRF

Fonte: Alex Gwo Jen Lan, 2022

Como mostrado na Tabela 3, os estudos são dedicados exclusivamente ao idioma inglês. Além disso, o gênero textual predominante é o *Twitter*. Em relação às representações textuais, é observado o uso frequente de modelos baseados em frequência de palavras como Análise Semântica Latente (LSA) e *Bag-Of-Words* (BoW). E no que se refere aos métodos, foram utilizados uma variedade de modelos como *Naive Bayes* (NB), regressão logística, *Long Short Term Memory* (LSTM) e *hinge-loss Markov random field* (HLMRF).

4 Materiais e métodos

O presente trabalho propõe o desenvolvimento de modelos de classificação de fundamentos morais a partir de texto baseados em modelos de língua pré-treinados de representação contextual, de modo a obter resultados superiores aos de abordagens tradicionais. De forma mais específica são abordadas duas formulações deste problema sob a perspectiva do Processamento de Língua Natural (PLN): a *classificação de fundamentos morais impessoais*, ou CFMI, e *classificação de fundamentos morais pessoais*, ou CFMP.

A tarefa de CFMI é tradicionalmente retratada na literatura como aquela que prediz fundamentos morais a partir de um texto. O caráter "impessoal" está associado ao fato de que essa tarefa avalia o texto considerando apenas a informação contida nele, de forma independente da moralidade do indivíduo que o escreveu. Dessa forma, textos semanticamente parecidos (e.g., opinião similares para um determinado assunto) estariam ligados aos mesmos fundamentos, e possivelmente classificados da mesma forma.

A tarefa de CFMP é inédita na literatura de PLN e consiste na predição de fundamentos morais associadas ao *indivíduo* que escreveu um determinado texto. O caráter "pessoal" refere-se ao fato de que a tarefa avalia textos com base não só nas informações presentes nele, mas também nos rótulos que expressam a moralidade do indivíduo, obtidas mediante um instrumento adequado, como o questionário de fundamentos morais em Silvino *et al.* (2016).

A principal diferença entre as duas tarefas está na abordagem do problema, que reside sobretudo na origem dos rótulos de classe. Em CFMI, o modelo proposto está classificando os textos de forma similar a problemas de análise de sentimentos da literatura (ZHANG; WANG; LIU, 2018), considerando rótulos de classe dependentes diretamente do significado comunicado. Em CFMP, o modelo pode ser visto como uma instância do problema de caracterização autoral (RANGEL; ROSSO, 2019) e está considerando textos de opinião com rótulos de classe dependentes do perfil moral do autor.

Embora similares do ponto de vista da aplicação, CFMI e CFMP são maneiras distintas de modelar fundamentos morais e podem inclusive resultar em classficações distintas para o mesmo texto de entrada. Para ilustrar esta questão, considera-se o exemplo extraído do córpus BRMoral (PAVAN *et al.*, 2020), que poderia ser considerado como Justiça, para CFMI, mas é definido como Pureza, para CFMP.

Exemplo: O fato é que por um grande período de tempo um grupo de pessoas foram menos favorecida, agora nada mais justo oferecer um meio para os mais esforçados ingressarem nas universidades.

Neste capítulo, são introduzidas as questões de pesquisa associadas às tarefas de CFMI e CFMP a serem investigadas neste estudo (seção 4.1). A seguir, também são apresentados os materiais utilizados que correspondem aos modelos de classificação (seção 4.2) e os conjuntos de dados (seção 4.3). Por fim, é descrito o procedimento adotado para a realização dos experimentos (seção 4.4).

4.1 Questões de pesquisa

As hipóteses gerais enunciadas no capítulo 1 propõem que os modelos de língua pré-treinados permitiriam classificar fundamentos morais a partir de textos com resultados superiores aos de abordagens tradicionais para CFMI e CFMP. Além disso, no caso do problema de CFMP, pode-se considerar ainda uma variante desta formulação que utiliza individualmente um dos tópicos do córpus BRmoral (que trata de oito tópicos de natureza moral organizados em subcórpus distintos), com o intuito de verificar a possível influência do conteúdo abordado em cada tópico sobre a tarefa de classificação. É importante mencionar que o tópico único corresponde ao tópico de maior correlação com o fundamento moral a ser classificado, conforme será detalhado na seção 4.4. Estas observações motivam as seguintes questões de pesquisa:

Q1: Modelos de língua pré-treinados aplicados à classificação de fundamentos morais expressos em textos, ou CFMI, alcançam resultados superiores às alternativas de *baseline*?

Q2: Modelos de língua pré-treinados aplicados à classificação de fundamentos morais de um indivíduo com base em textos multitópicos de sua autoria, ou CFMP, alcançam resultados superiores às alternativas de *baseline*?

Q3: Modelos de língua pré-treinados aplicados à classificação de fundamentos morais de um indivíduo com base em textos de tópico único de sua autoria, ou CFMP, alcançam resultados superiores às alternativas de *baseline*?

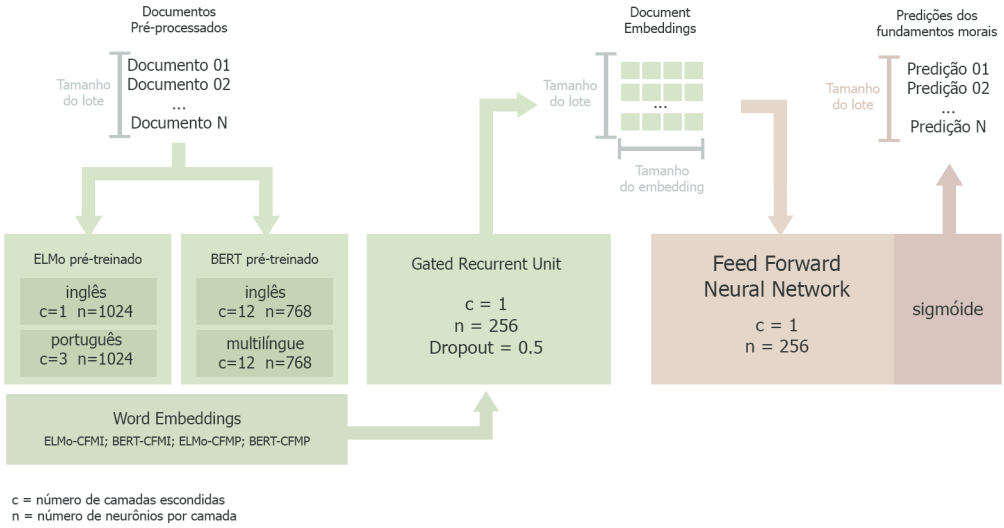
Essas questões serão verificadas comparando-se as estratégias propostas, desenvolvidas com uso de modelos de língua pré-treinados do tipo ELMo (PETERS *et al.*, 2018) e BERT (DEVLIN *et al.*, 2019), com os sistemas de *baseline* definidos a partir de estudos já existentes quando possível ou modelos baseados em representações textuais tradicionais. Para a avaliação, será utilizada uma das métricas mais aplicadas em tarefas de aprendizado de máquina, a F1 macro.

As abordagens de CFMI e CFMP usadas para responder as questões de pesquisa foram definidas como tarefas de classificação binária nas quais os modelos recebem um texto como entrada (e.g., *tweets* ou texto de opinião) e predizem os fundamentos morais expressos no texto, ou os fundamento morais associados ao indivíduo (autor) que o produziu. Para cada tipo de abordagem, temos cinco tarefas de classificação binária independentes, cada uma representando a presença ou ausência de um fundamento moral.

4.2 Modelos de classificação

No contexto do presente estudo, as tarefas de CFMI e CFMP são aplicadas respectivamente para os idiomas em inglês e português brasileiro. Essa configuração se deve essencialmente à disponibilidade do conjunto de dados para cada uma dessas tarefas na literatura. Para cada tarefa, construíram-se modelos baseados em *embeddings* contextuais pré-treinados. Dessa forma, temos quatro tipos de modelos: ELMo-CFMI e BERT-CFMI, para o inglês, além do ELMo-CFMP e BERT-CFMP, para o português brasileiro. É importante mencionar que os modelos propostos seguem a mesma estrutura geral, como ilustrado na Figura 7.

Figura 7 – Arquitetura do modelo proposto



Fonte: Alex Gwo Jen Lan, 2022

Como passo inicial dentro da estrutura, são gerados os *embeddings* das palavras de um documento pré-processado a partir de modelos pré-treinados de língua (ELMo ou BERT). A implementação ELMo para a versão em inglês (PETERS *et al.*, 2018) consiste numa única camada escondida de tamanho 1.024, abrangendo 13.6 milhões de parâmetros, enquanto que a versão para o português (CASTRO, 2019) é constituída por três camadas escondidas de tamanho 1.024. A implementação BERT para a versões em inglês e multilíngue (DEVLIN *et al.*, 2019) é composta por 12 camadas escondidas de tamanho 768 e 12 cabeças de auto-atenção, compreendendo, respectivamente, 110 milhões e 179 milhões de parâmetros.

Em seguida, os *embeddings* das palavras gerados pelos modelos pré-treinados são utilizados para criar a representação do documento analisado. Para isso, é aplicado um modelo GRU composto por uma camada escondida com 256 neurônios e *dropout* em 0.5. Depois, o *document embedding* criado é passado para a última camada da estrutura.

A última camada é uma *Feed Forward Neural Network* (FFNN) de tamanho 256. Nesta rede, é calculada a equação linear ($s = Wx$), onde x é um *document embedding*, W,o peso e s, a saída. Por fim, o valor resultante é processado por uma função sigmóide para gerar a predição de um fundamento. O treinamento foi realizado em 10 épocas, com tamanho de lote igual a 32, taxa de aprendizado em 0.1 e entropia cruzada como função de perda. Essas configurações de implementação, realizadas com a biblioteca *Flair*, são baseadas no estudo apresentado em (AKBIK *et al.*, 2019).

4.2.1 Modelos baseados em *embeddings* dependentes de contexto

Os modelos propostos ELMo-CFMI e ELMo-CFMP são construídos com base em modelos de língua pré-treinados de *Embeddings for Language Model* (ELMo) e ajustados para a tarefa de classificação de fundamentos morais a partir de texto. No caso do ELMo-CFMI, o ajuste foi realizado a partir da implementação ELMo para o inglês (PETERS *et al.*, 2018), enquanto que o ELMo-CFMP foi ajustado para a versão em português brasileiro (CASTRO, 2019).

Os modelos propostos BERT-CFMI e BERT-CFMP consideram em sua composição modelos pré-treinados de *Bidirectional Encoder Representations from Transformers* (BERT) que foram ajustados para a tarefa de classificação de cada um dos fundamentos morais. Para BERT-CFMI, utilizou-se o modelo BERT para o inglês e para o BERT-CFMP, considerou-se o modelo BERT multilíngue (DEVLIN *et al.*, 2019). A escolha pela versão multilíngue em detrimento de um modelo BERT treinado para o português brasileiro (SOUZA; NOGUEIRA; LOTUFO, 2020) foi devida aos resultados superiores alcançados pelo primeiro durante os testes, com uma diferença geral de 10% nos valores de F1 macro.

4.2.2 Modelos *baseline* para CFMI

Como sistemas *baseline* para CFMI, foram considerados os modelos utilizados em Araque, Gatti e Kalimeri (2020). Os mesmos se encontram disponíveis no repositório de código dos próprios autores[1]. Este trabalho foi escolhido como referência para a comparação dos resultados do modelo proposto por utilizar, de forma exclusiva na literatura, o conjunto de dados MFTC (HOOVER *et al.*, 2020). Além disso, também foi elaborado um modelo de regressão logística de n-gramas de caracteres, aqui denominado *reglog.char*, para ser usado como um *baseline* possivelmente mais robusto desse tipo de tarefa. Abaixo, estão listadas as abordagens de extração de características presente em todos estes modelos.

- O modelo de contagem (ou *count*), apresentado em Araque, Gatti e Kalimeri (2020), emprega o método descrito na seção 3.1 dos trabalhos correlatos. De forma sucinta, o modelo define a classe de um documento com base na presença majoritária das

[1] github.com/oaraque/moral-foundations

palavras associadas a um fundamento moral específico. Essas palavras são as mesmas presentes no DFM original (GRAHAM; HAIDT; NOSEK, 2009).

- Moral Freq (ou *freq*), proposto em Araque, Gatti e Kalimeri (2020), representa um documento de texto considerando a quantidade de palavras que expressam um fundamento moral específico. Para definir se uma palavra tem um determinado traço moral, verifica-se se o seu valor de valência moral na versão expandida de DFM, o *MoralStrength*, está acima de um determinado limiar, que segundo os autores do estudo em que foi proposto, por conta de propriedades de generalização do léxico moral, é 5. Realizou-se uma adaptação nesse modelo para considerar apenas o fundamento em si, ao invés de suas subdivisões de vício e virtude. Assim, este modelo produz como saída um vetor de representação de 5 dimensões, contendo as frequências para cada fundamento.
- O modelo de unigrama e as suas variantes, definidos em Araque, Gatti e Kalimeri (2020), tratam daqueles que utilizam uma representação do tipo BoW para o documento. As variantes consistem na combinação dos outros dois métodos previamente mencionados. Dessa forma, temos a utilização dos modelos de unigrama, de unigrama + *count*, de unigrama + *freq* e de unigrama + *count* + *freq*. A saída destes modelos é a concatenação do vetor produzido por cada um dos componentes da combinação.
- A *reglog.char*, construído no presente trabalho, é um modelo baseado em regressão logística sobre contagens TF-IDF, reduzido às suas k melhores características por meio de uma seleção de atributos univariada a partir das pontuações calculadas pela função F-valor da análise de variância (ANOVA). Estas características são representadas no nível de n-grams, com n entre 2 e 8. Este intervalo foi definido empiricamente como aquele que forneceu os melhores resultados durante os testes com o córpus.

Ao todo, foram avaliados sete modelos como *baseline* de CFMI, correspondendo aos modelos *count*, *freq*, as quatro versões de unigrama e o *reglog.char*. Para todas estas abordagens, foram utilizados como modelo de classificação a regressão logística treinada com validação cruzada em 10 partições. Nota-se a ausência dos modelos de *Moral Stats* e *SIMON*, que estão descrito na subseção 3.2.9 dos trabalhos correlatos, uma vez que estes modelos não estavam disponíveis no repositório de código dos autores e o estudo em questão não apresentava detalhamento suficiente para sua reimplementação.

4.2.3 Modelos *baseline* para CFMP

Como sistemas *baseline* para CFMP, foram construídos dois modelos de regressão logística (*reglog*). Este tipo de modelo têm apresentado resultados superiores para tarefas de classificação de texto (GARTEN *et al.*, 2016). A implementação foi realizada com a biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011). Os modelos são detalhados como segue.

- *reglog.word* usa a representação de frequência de palavras TF-IDF a partir dos k termos com as melhores pontuações. As k palavras são selecionadas do conjunto de treinamento por meio do cômputo do F-valor da análise de variância (ANOVA).
- *reglog.char* é o mesmo modelo descrito na subseção 4.2.2.

4.3 Conjunto de dados

Os córpus selecionados para as tarefas de CFMI em inglês e de CFMP em português brasileiro correspondem, respectivamente, ao *Moral Foundations Twitter Corpus* (MFTC) e ao BRMoral, cujas configurações originais estão descritas na seção 2.2. Para se adequar às tarefas de classificação de texto propostas, ambos conjuntos de dados foram submetidos a um pré-processamento, resultando na normalização dos textos e na remoção de instâncias vazias. Na Tabela 4, são sumarizadas as estatísticas descritivas dos conjuntos de dados efetivamente utilizados nos experimentos descritos a seguir.

Tabela 4 – Estatísticas descritivas dos córpus utilizados

Córpus	Idioma	Domínio	Instâncias	Palavras	Tipos
MFTC	IN	*tweets*	34.922	516.349	36.712
BRMoral (all)	PT-BR	opinião	3.976	214.169	11.480
BRMoral (aborto)	PT-BR	opinião	497	30.790	3.376
BRMoral (casamento gay)	PT-BR	opinião	497	23.321	2.679
BRMoral (controle de armas)	PT-BR	opinião	497	27.645	3.493
BRMoral (cotas raciais)	PT-BR	opinião	497	29.028	3.473
BRMoral (isenção fiscal de igrejas)	PT-BR	opinião	497	22.720	3.138
BRMoral (legalização das drogas)	PT-BR	opinião	497	27.129	3.428
BRMoral (maioridade penal)	PT-BR	opinião	497	26.769	3.304
BRMoral (pena de morte)	PT-BR	opinião	497	26.767	3.567

Fonte: Alex Gwo Jen Lan, 2022

Como pode ser observado na Tabela 4, a utilização do MFTC considerou todos os setes domínios abrangidos no conjunto de dados original (i.e., 2016 *Presidential Election*,

All Lives Matter, *Baltimore Protest*, *Black Lives Matter* e etc.). A distribuição das classes desta versão do MFTC pode ser visualizada na Tabela 5.

Tabela 5 – Distribuição de classes no MFTC

Classe	Instâncias
Cuidado	5.538
Justiça	4.756
Lealdade	4.139
Autoridade	2.577
Pureza	2.564

Fonte: Alex Gwo Jen Lan, 2022

Além da distribuição mostrada na Tabela 5, no córpus há ainda 15.348 instâncias de texto rotulado como "não-moral". Este conjunto, assim como em Araque, Gatti e Kalimeri (2020), é utilizado para formar o conjunto de instâncias negativas de cada fundamento moral. De forma mais específica, para cada fundamento moral é selecionado aleatoriamente um conjunto de instâncias não-morais de mesmo tamanho, formando assim problemas perfeitamente balanceados. É importante destacar que as classes positivas referem-se à presença do dado fundamento num documento, tanto em seu aspecto de virtude quanto o de vício, e as classes negativas referem-se à ausência deste fundamento.

No caso do córpus BRmoral, também é possível notar que a sua aplicação para cada um dos experimentos é realizada sob diferentes configurações. Na versão BRMoral(all), considera-se o texto de todos os 8 tópicos contidos no conjunto original, totalizando 8 córpus * 497 instâncias = 3.976 instâncias. As outras variações do BRMoral, referidas também como BRMoral(ind), contêm o texto de apenas um dos tópicos.

Considerando-se que a rotulação de classes ternária ("*Low*", "*Avg*"e "*High*") originalmente disponibilizada pelo córpus BRmoral tornava certos fundamentos morais muito esparsos para o propósito de classificação, este esquema foi substituído por uma nova rotulação binária ("*Low*"e "*High*"). De forma mais específica, partindo-se dos escores de fundamentos morais também disponibilizados pelo córpus em uma escala de 0 a 30 (e que são obtidas a partir das respostas ao QFM apresentadas pelos indivíduos participantes do córpus, conforme discutido na subseção 2.2.2), é gerada a mediana que servirá como um limiar para decidir à qual classe uma determinada instância pertence para cada fundamento. Caso a pontuação de uma dada instância esteja abaixo do limiar, a classe seria "*Low*", caso contrário, se estiver acima, seria "*High*". A distribuição dessas duas classes no BRMoral(all) e nas versões com apenas um córpus é mostrada na Tabela 6.

Tabela 6 – Distribuição de classes nas configurações do BRMoral

	BRMoral (all)		BRMoral (ind)	
Classe	Low	High	Low	High
Cuidado	2.264	1.712	283	214
Justiça	2.208	1.768	276	221
Lealdade	2.120	1.856	265	232
Autoridade	2.224	1.752	278	219
Pureza	2.024	1.952	253	244

Fonte: Alex Gwo Jen Lan, 2022

Nos experimentos de CFMI, é usada a versão do MFTC descrita anteriormente. Este conjunto de dados se caracteriza por ser uma coleção de *tweets* no idioma em inglês. A distribuição dos dados em conjuntos de treinamento, validação e teste deste córpus pode ser visualizada na Tabela 7.

Tabela 7 – Distribuição de instâncias por classe em treinamento, validação e teste (MFTC)

Classe	Treinamento	Validação	Teste
Cuidado	7.974	886	2.216
Justiça	6.848	761	1.903
Lealdade	5.959	663	1.656
Autoridade	3.710	413	1.031
Pureza	3.691	411	1.026

Fonte: Alex Gwo Jen Lan, 2022

Nos experimentos de CFMP, são usadas as nove versões do BRMoral apresentadas previamente. Estes conjuntos de dados são constituídos por uma série de textos de opinião no idioma em português brasileiro. A distribuição dos dados em conjuntos de treinamento, validação e teste do BRMoral nas duas formulações de CFMP é mostrada na Tabela 8.

Tabela 8 – Distribuição de instâncias em treinamento, validação e teste (BRMoral)

Córpus	Treinamento	Validação	Teste
BRMoral(all)	2.862	318	796
BRMoral(ind)	357	40	100

Fonte: Alex Gwo Jen Lan, 2022

Como mostrado na Tabela 8, BRMoral(all) e BRMoral(ind) possuem a mesma distribuição de instâncias em treinamento, validação e teste para todos os fundamentos. Isto se deve ao fato de que o córpus BRmoral apresenta a mesma quantidade de instâncias para estes traços morais, variando apenas no número de classes positivas e negativas.

Na formulação de CFMP de tópico único, a escolha do tópico ótimo para cada fundamento moral foi realizada utilizando classificador de *reglog.word*, o mesmo descrito na subseção 4.2.3, para avaliar todas aquelas oito versões do BRMoral. A decisão em executar *reglog.word* nesta etapa foi devida ao custo de tempo e recurso computacional para desempenhar a grande quantidade de execuções (8 tópicos * 5 fundamentos = 40 execuções). Os resultados em F1 macro desta análise estão descritos na Tabela 9.

Tabela 9 – Resultados *reglog.word* fundamentos versus tópicos (F1 macro)

Tópico	Cuidado	Justiça	Lealdade	Autoridade	Pureza
aborto	0.50	0.36	0.35	0.57	**0.61**
casamento *gay*	0.50	**0.53**	0.56	0.56	0.55
controle de armas	0.35	0.36	**0.58**	0.36	0.34
cotas raciais	0.36	0.36	0.49	0.61	0.55
isenção fiscal de igrejas	0.36	0.37	0.56	0.51	0.58
legalização drogas	0.36	0.51	0.53	0.53	0.56
maioridade penal	0.50	0.36	0.47	**0.63**	0.58
pena de morte	**0.55**	0.36	0.51	0.58	0.53

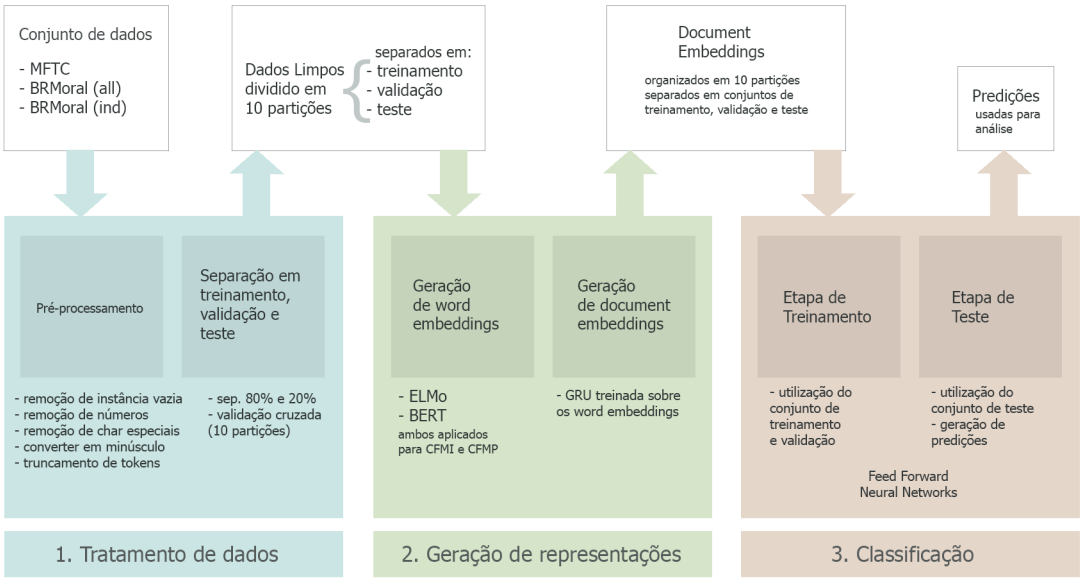
Fonte: Alex Gwo Jen Lan, 2022

Conforme a Tabela 9, os tópicos ótimos de cada tarefa de classificação foram definidos como pena de morte para Cuidado, casamento *gay* para Justiça, controle de armas para Lealdade, maioridade penal para Autoridade e aborto para Pureza. Estes subconjuntos do BRmoral - e não o córpus inteiro, que é usado para investigar a questão de pesquisa Q2 - serão utilizados para investigar a questão de pesquisa Q3.

4.4 Procedimentos

Os experimentos conduzidos para investigar as questões de pesquisa Q1, Q2 e Q3 enunciados na seção 4.1 seguem um mesmo fluxo geral ilustrado na Figura 8. Como observado, este fluxo é definido em etapas de tratamento de texto, geração de *embeddings* e classificação, representados respectivamente como retângulos azul, verde e marrom na imagem. Em cada uma dessas etapas, é utilizado o produto gerado pelo passo anterior como entrada do procedimento atual (retângulos brancos).

Figura 8 – Fluxo geral da aplicação do modelo proposto



Fonte: Alex Gwo Jen Lan, 2022

Como etapa inicial, realizou-se um tratamento de todos os conjuntos de dados. Este procedimento é essencial para a remoção de ruídos que possam prejudicar o desempenho dos modelos. Para isso, o processo de preparação dos dados consistiu em primeiro remover instâncias vazias, ou seja, aquelas que contém unicamente espaços vazios ou "*na*". Em seguida, os textos foram pré-processados e organizados em conjuntos de treinamento, validação e teste.

O pré-processamento para o córpus em português brasileiro inclui etapas de conversão dos caracteres em minúsculos e remoção de caracteres numéricos e excesso de espaços em brancos. Para o córpus em inglês, além das etapas mencionadas anteriormente, por se tratar de textos produzidos em redes sociais, foram incluídas passos adicionais como a remoção de caracteres especiais (i.e. *hashtags* e menções com @), *urls* e *links*. Além disso, como etapa adicional para as tarefas com BERT, visando o funcionamento deste tipo de arquitetura, considerou-se o truncamento dos textos para limitá-los a 512 *tokens*.

No próximo passo, o texto pré-processado, tanto para o português quanto para o inglês, é dividido em conjunto de treinamento-validação e teste, que correspondem respectivamente a 80% e 20% do conjunto de dados. Depois, o conjunto de treinamento-validação é submetido à etapa de validação cruzada (*k-fold*). Neste momento, o conjunto de dados é dividido em 10 compartimentos (k=10).

Para fins de avaliação, foi utilizada a métrica F1 macro. Esta métrica foi escolhida devida à sua ampla utilização nos trabalhos correlatos (seção 3), especialmente naqueles aos quais foram utilizados como *baseline* do presente estudo. O cálculo da medida F1 simples considera a combinação das métricas de precisão e revocação (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006), como mostrado abaixo:

$$F1 = 2 * \frac{precisão * revocação}{precisão + revocação}$$

A partir da fórmula apresentada, é calculada a métrica final de avaliação F1 macro. Primeiramente, define-se o valor da medida F1 para cada uma das classes da tarefa e, por conseguinte, realiza-se a média não ponderada desses valores individuais para se obter a métrica final de F1 macro.

Como forma de avaliar a significância estatística destes resultados, foi utilizado o teste de McNemar (MCNEMAR, 1947). Este teste estatístico descreve o quanto dois modelos de classificação binária concordam ou discordam. Para isso, toma-se como entrada o número de predições corretas e incorretas de cada classificador em uma tabela de contingência de dimensão 2. O cálculo do teste de McNemar é mostrado na equação abaixo (MCNEMAR, 1947).

$$x^2 = \frac{(D - A)^2}{D + A}$$

Nesta equação, x^2 é o teste estatístico, D corresponde ao número de predições que um dado classificador 1 acertou e o classificador 2, errou, e A, o número de predições que o classificador 2 acertou, e o 1, errou. Neste cálculo, é possível observar que o teste de McNemar captura a quantidade de erros e acertos nas predições de dois modelos considerados na comparação.

Se as quantidades de acertos e erros entre os dois classificadores são similares, é concluído que ambos modelos apresentam erros na mesma proporção, apenas em instâncias diferentes no conjunto de teste e assim, o resultado da estatística pode não ser significativa. Caso contrário, se as quantidades não são similares. Isso significa que ambos modelos não apenas produzem erros diferentes, mas de fato diferem na proporção relativa de erros no conjunto de teste e, dessa forma, podem ser considerados como tendo resultados estatisticamente distintos.

5 Resultados

Este capítulo apresenta os resultados de três experimentos que objetivam responder as questões de pesquisa Q1, Q2 e Q3 enunciadas na seção 4.1. Os resultados apresentados ao longo deste capítulo são reportados na forma de medida F1 macro, e diferenças entre as alternativas consideradas são avaliadas com uso do teste de McNemar (MCNEMAR, 1947), conforme discutido na seção anterior.

Em todos os resultados apresentados, o modelo de melhor medida F1 macro para cada classe é destacado em negrito, e a significância estatística da diferença entre esta e a segunda melhor alternativa é assinalada como * para $p < 0{,}05$, ** para $p < 0{,}01$ e *** para $p < 0{,}001$. Além disso, os modelos foram organizados em grupos estatisticamente distintos considerando um p mínimo $< 0{,}05$. A seguir, nas seções de 5.1 a 5.3 são discorridos os resultados de cada um dos experimentos. Por fim, na seção 5.4, estes resultados são discutidos e complementados.

5.1 Questão Q1: Classificação de fundamentos morais impessoal (CFMI) em inglês

Com o intuito de avaliar a questão de pesquisa Q1, conduziu-se o experimento de CFMI usando o córpus MFTC em inglês. Os resultados são apresentados na Tabela 10.

Tabela 10 – Resultados do experimento de CFMI em inglês para MFTC (F1 macro)

Estratégia	Cuidado	Justiça	Lealdade	Autoridade	Pureza
count	0.66	0.36	0.51	0.32	0.36
freq	0.74	0.72	0.71	0.69	0.72
unigrama	0.56	0.52	0.51	0.57	0.58
unigrama + count	0.60	0.53	0.53	0.60	0.60
unigrama + freq	0.64	0.63	0.65	0.70	0.62
unigrama + count + freq	0.63	0.63	0.66	0.69	0.62
reglog.char	0.88	0.89	**0.86**	**0.89**	**0.85**
ELMo-CFMI	0.87	0.90	0.85	0.87	0.84
BERT-CFMI	**0.89**	**0.91**	**0.86**	0.87	**0.85**

Fonte: Alex Gwo Jen Lan, 2022

Como mostrado na Tabela 10, os valores de F1 macro sugerem como possível resposta para a questão de pesquisa Q1 que o uso de modelos pré-treinados de língua são na maioria dos casos superiores às alternativas, embora em alguns casos por pequena margem (especialmente em comparação ao modelo *reglog.char*). Os modelos BERT e ELMo

propostos superam todos os sistemas de *baseline* de Araque, Gatti e Kalimeri (2020) (ou seja, os modelos count, freq, unigrama, unigrama + count, unigrama + freq e unigrama + count + freq) por ampla margem.

Os modelos considerados foram ordenados em grupos estatisticamente similares de acordo com o teste de McNemar para cada classe de fundamento moral. Os grupos homogêneos obtidos podem ser visualizados nas Tabelas 11, 12, 13, 14 e 15.

Tabela 11 – Grupos homogêneos para CFMI (Cuidado)

Modelos	F1 macro	Grupos				
BERT-CFMI	**0.89**	A				
reglog.char	0.88	A				
ELMo-CFMI	0.87	A				
freq	0.74		B			
count	0.66			C		
unigram + freq	0.64			C		
unigram+count+freq	0.63			C		
unigram+count	0.60				D	
unigram	0.56					E

Fonte: Alex Gwo Jen Lan, 2022

Tabela 12 – Grupos homogêneos para CFMI (Justiça)

Modelos	F1 macro	Grupos				
BERT-CFMI	**0.91**	A				
ELMo-CFMI	0.90	A				
reglog.char	0.89	A				
freq	0.72		B			
unigram+count+freq	0.63			C		
unigram+freq	0.63			C		
unigram+count	0.53				D	
unigram	0.52				D	
count	0.36					E

Fonte: Alex Gwo Jen Lan, 2022

Tabela 13 – Grupos homogêneos para CFMI (Lealdade)

Modelos	F1 macro	Grupos			
BERT-CFMI	**0.86**	A			
reglog.char	0.86	A			
ELMo-CFMI	0.85	A			
freq	0.71		B		
unigram+count+freq	0.66			C	
unigram+freq	0.65			C	
unigram+count	0.53				D
unigram	0.51				D
count	0.51				D

Fonte: Alex Gwo Jen Lan, 2022

Tabela 14 – Grupos homogêneos para CFMI (Autoridade)

Modelos	F1 macro	Grupos			
reglog.char	**0.89**	A			
BERT-CFMI	0.87	A			
ELMo-CFMI	0.87	A			
unigram+freq	0.70		B		
freq	0.69		B		
unigram+count+freq	0.69		B		
unigram+count	0.60			C	
unigram	0.57			C	
count	0.32				D

Fonte: Alex Gwo Jen Lan, 2022

Tabela 15 – Grupos homogêneos para CFMI (Pureza)

Modelos	F1 macro	Grupos				
BERT-CFMI	**0.85**	A				
reglog.char	0.85	A				
ELMo-CFMI	0.84	A				
freq	0.72		B			
unigram+count+freq	0.62			C		
unigram+freq	0.62			C		
unigram+count	0.60			C		
unigram	0.58				D	
count	0.36					E

Fonte: Alex Gwo Jen Lan, 2022

Como ilustrado nas Tabelas 11 a 15, os modelos de BERT-CFMI e ELMo-CFMI sempre fazem parte do grupo A, pois possuem os melhores resultados e a mesma proporção de erros entre si para todos os fundamentos.

5.2 Questão Q2: Classificação de fundamentos morais pessoais (CFMP) multitópico em português

Com o intuito de avaliar a questão de pesquisa Q2, conduziu-se o experimento de CFMP multitópico usando o córpus BRMoral (all) em português brasileiro. Os resultados são apresentados na Tabela 16.

Tabela 16 – Resultados do experimento de CFMP multitópico em português brasileiro para BRMoral(all) (F1 macro)

Estratégia	Cuidado	Justiça	Lealdade	Autoridade	Pureza
reglog.word	0.47	0.47	0.56	0.55	0.58
reglog.char	***0.79**	***0.80**	***0.78**	***0.77**	***0.81**
ELMo-CFMP	0.53	0.51	0.52	0.52	0.53
BERT-CFMP	0.53	0.47	0.52	0.52	0.41

Fonte: Alex Gwo Jen Lan, 2022

Como observado na Tabela 16, os resultados sugerem como possível resposta para a questão de pesquisa Q2 que o uso de modelos pré-treinados de língua não alcançam resultados superiores às alternativas *baseline*, com base em textos multitópicos. Nota-se também que o modelo de *reglog.char* se sobressaiu em relação aos demais. Uma possível explicação para os resultados pode estar relacionada à natureza do córpus BRmoral, que é formado de textos longos que precisam ser truncados para uso pelos modelos neurais. Outras considerações sobre este resultado são discutidas na seção 5.4. Ademais, em vista da diferença expressiva entre *reglog.char* e as demais alternativas, por simplicidade omitimos a apresentação da divisão em grupos homogêneos, sendo possível intuir que *reglog.char* formaria um grupo (A) distinto dos demais.

5.3 Questão Q3: Classificação de fundamentos morais pessoais (CFMP) de tópico único em português

Com o intuito de avaliar a questão de pesquisa Q3, conduziu-se o experimento de CFMP de tópico único usando cinco versões do córpus BRMoral em português brasileiro. De forma específica, estas cinco versões foram selecionadas de acordo com a melhor correspondência para cada uma das classes de fundamentos, conforme descrito na seção 4.3. Os resultados são apresentados na Tabela 17.

Tabela 17 – Resultados do experimento de CFMP de tópico único em português brasileiro para BRMoral(ind) (F1 macro)

Estratégia	Cuidado pena.morte	Justiça casam.*gay*	Lealdade contr.armas	Autoridade maior.penal	Pureza aborto
reglog.word	0.55	0.53	0.58	0.63	0.61
reglog.char	****0.79**	*****0.79**	*****0.85**	***0.83**	****0.84**
ELMo-CFMP	0.53	0.54	0.57	0.51	0.55
BERT-CFMP	0.51	0.50	0.49	0.50	0.50

Fonte: Alex Gwo Jen Lan, 2022

No caso da tarefa de CFMP de tópico único, os resultados da Tabela 17 são similares aos do experimento anterior (CFMP multitópico). Novamente, os modelos pré-treinados apresentaram desempenho inferior ao do modelo *reglog.char*. Assim a resposta para a questão Q3 é novamente negativa, ou seja, modelos pré-treinados de língua não alcançam um desempenho superior às alternativas *baselines*, com base em textos de tópico único.

Os modelos considerados foram ordenados em grupos estatisticamente similares de acordo com o teste de McNemar para cada classe de fundamento moral. Estes grupos homogêneos podem ser visualizados nas Tabelas 18, 19, 20, 21 e 22.

Tabela 18 – Grupos homogêneos para CFMP de tópico único (Cuidado - pena.morte)

Modelos	F1 macro	Grupos
reglog.char	**0.79**	A
reglog.word	0.55	B
ELMo-CFMP	0.53	B
BERT-CFMP	0.51	B

Fonte: Alex Gwo Jen Lan, 2022

Tabela 19 – Grupos homogêneos para CFMP de tópico único (Justiça - casam.*gay*)

Modelos	F1 macro	Grupos
reglog.char	**0.79**	A
reglog.word	0.54	B
ELMo-CFMP	0.53	B
BERT-CFMP	0.50	B

Fonte: Alex Gwo Jen Lan, 2022

Tabela 20 – Grupos homogêneos para CFMP de tópico único (Lealdade - contr.armas)

Modelos	F1 macro	Grupos	
reglog.char	**0.85**	A	
reglog.word	0.58		B
ELMo-CFMP	0.57		B
BERT-CFMP	0.49		B

Fonte: Alex Gwo Jen Lan, 2022

Tabela 21 – Grupos homogêneos para CFMP de tópico único (Autoridade - maior.penal)

Modelos	F1 macro	Grupos	
reglog.char	**0.83**	A	
reglog.word	0.63		B
ELMo-CFMP	0.51		B
BERT-CFMP	0.50		B

Fonte: Alex Gwo Jen Lan, 2022

Tabela 22 – Grupos homogêneos para CFMP de tópico único (Pureza - aborto)

Modelos	F1 macro	Grupos	
reglog.char	**0.84**	A	
reglog.word	0.61		B
ELMo-CFMP	0.55		B
BERT-CFMP	0.50		B

Fonte: Alex Gwo Jen Lan, 2022

Os grupos visualizados nas Tabelas de 18 a 22 indicam que *reglog.char*, por possuir um desempenho bem distinto para todos os fundamentos em relação aos outros, representa sozinho o grupo A, enquanto que os modelos BERT-CFMP e ELMo-CFMP se encontram no mesmo grupo B, junto com o *reglog.word*.

Como indicado nesta seção, os resultados são similares ao do experimento anterior. A redução da tarefa a um tópico específico não melhorou os resultados dos modelos BERT-CFMP e ELMo-CFMP, mas diminuiu a acurácia média do modelo *reglog.char*, possivelmente em virtude do menor volume de dados disponível (no presente caso utilizando apenas um subconjunto de textos ao invés dos oito conjuntos concatenados, como descrito na seção 5.2).

5.4 Considerações

Este capítulo relatou os resultados de três experimentos abordando as tarefas de CFMI, CFMP multitópico e CFMP de tópico único. No primeiro experimento, foram empregados modelos propostos de BERT-CFMI e ELMo-CFMI sobre o córpus MFTC no idioma inglês. Para o segundo e terceiro experimentos, aplicou-se os modelos BERT-CFMP e ELMo-CFMP sobre o córpus BRMoral no idioma português brasileiro.

Em relação à questão de pesquisa Q1, o experimento de CFMI sugere que os modelos de língua pré-treinados de *embeddings* contextuais alcançam na maioria dos casos resultados superiores às alternativas *baselines*. Em especial, o modelo BERT-CFMI apresentou os maiores valores de F1 macro, com pequenas diferenças quando comparados com os de ELMo-CFMI.

Em relação à questão de pesquisa Q2, o experimento de CFMP multitópico sugere que os modelos de língua pré-treinados de *embeddings* contextuais não alcançam resultados superiores às alternativas *baselines*, com base em textos multitópicos. Neste caso, os modelos BERT-CFMP e ELMo-CFMP apresentaram um desempenho próximo ao modelo de *reglog.word*.

Em relação à questão de pesquisa Q3, o experimento de CFMP de tópico único sugere que os modelos de língua pré-treinados de *embeddings* contextuais não alcançam resultados superiores às alternativas *baselines*, com base em textos de tópico único. Como no experimento anterior, que também considera o córpus BRMoral, os modelos BERT-CFMP e ELMo-CFMP geraram resultados similares ao modelo de *reglog.word*.

Um ponto importante a se notar em todos os experimentos foram os resultados obtidos a partir do modelo de *reglog.char*. Esta abordagem, construída inicialmente como um *baseline* mais robusto para ambas tarefas, alcançou valores de F1 macro próximos aos modelos de *embeddings* contextuais de CFMI, assim como se destacou em relação aos modelos de CFMP. No restante desta seção, são apresentadas algumas considerações adicionais sobre o desempenho dos classificadores desenvolvidos, divididos em duas partes. Na seção 5.4.1, são apresentadas as características consideradas mais relevantes para cada tarefa, e na seção 5.4.2, é discutida uma classe específica de exemplo de texto classificado incorretamente pelos modelos BERT e ELMo na tarefa de CFMP.

5.4.1 Relevância das características de aprendizado

Para compreender melhor o desempenho dos modelos de CFMI e CFMP, verificou-se as caraterísticas mais relevantes para cada fundamento moral. Para isso, considerou-se a utilização da biblioteca *eli5*[1] de explicação de modelos de aprendizado de máquina. A extração foi realizada a partir de *reglog.char* por ter o melhor desempenho dentre os modelos suportados pela biblioteca. As características mais relevantes para CFMI e CFMP podem ser respectivamente visualizadas nas Tabelas 23 e 24.

Como observado nas tabelas, as três colunas à esquerda contém informações referentes aos termos que mais influenciaram na predição de classes positivas, enquanto que as três colunas à direita possuem informações dos termos com maior influência na predição das classes negativas. Por ser um modelo de n-gramas de caracteres, as características selecionadas nem sempre possuem uma interpretação humana óbvia, e que por este motivo são apresentadas as características propriamente ditas e exemplos extraídos do córpus onde elas ocorrem.

[1] ⟨https://eli5.readthedocs.io/en/latest/⟩

Tabela 23 – Palavras com maior influência no MFTC em cada fundamento

Cuidado

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+6.843	a rapi	OMFG HE'S **A RAPI**ST	-3.919	a kille	MILITI**A KILLE**D ATTACKED
+6.798	ace if	the Storm We ll F**ace if** Romney	-3.813	again c	we meet **again c**an't wait
+6.381	a prc	Uighur Jihadis: they're not just **a PRC**	-3.777	am m	Mainstre**am M**edia not reporting
+6.289	allis	you ain't six bitch - **Alli**sons Mom	-3.718	affin	I have 150 day boosters on **affin**ity
+6.279	a fuc	I can never give **a fuc**k	-3.674	a day k	A pussy **a day k**eeps the devil away
+6.049	allies	defend themselves at his r**allies**	-3.275	absolut	**Absolut**e morons protesting
+6.011	am for	I **am for** peace!	-3.213	a murde	you're a coward and **a murde**rer
+5.886	a batt	Saddened to see ... into **a batt**le	-3.142	acros	extreme devastation **acros**s Bethesda
+5.868	a baby	punished for not having **a baby**	-3.139	a picks	Stay calm rock **a Picks** For Peace
+5.696	a skin	a sin problem, not **a skin** problem	-3.127	a cent	suffered terrorism for well over **a cent**ury

Justiça

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+21.941	a	-	-13.910	a fucki	GO AWAY, your **a fucki**ng fraud!
+17.579	airpor	MSP **airpor**t listening to white couple	-13.524	aka sa	hate Don Lemon cnn sandy **aka sa**lly
+15.700	a k	abused **a k**id ... Equal justice	-13.417	all i w	**all I w**ant is justice
+15.302	a metho	as **a method** of shock and awe	-13.091	amaze	You'd be **amaze**d at the racist
+14.950	acts op	unjust **acts op**pression racial hatred	-12.918	aeri	**Aeri**al view shows ... of protesters
+14.897	a high	**a high**er court than the courts of justice	-12.362	an und	in **an und**emocratic irreformable union
+14.155	a knee	When **a knee** is in your neck	-11.586	a quart	that bitch aint worth **a quart**er
+13.935	am cr	I **am cr**uel to myself	-11.222	alive	should be **alive**
+13.808	ade	Thank you just doesn't seem **ade**quate	-10.874	alex	**Alex** Jones is spreading lies
+13.299	aid sa	s**aid Sa**ndy ... punishment for Gay	-10.656	almos	**Almos**t. Welp... rooting for equality

Lealdade

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+6.885	aclu	corrupt **ACLU** protecting the racist	-4.243	amp k	if there is decency **amp k**indness left
+6.747	abt abu	**abt abu**se of power eg Trump!	-4.112	a thon	Every female owns a **thon**g
+6.032	act f	**Act f**or Humanity	-4.031	a cripp	he isn't **cripp**le like your hero Roach
+5.451	and say	can we be honest **and say** black lives	-4.025	a prot	Looting ... doesn't make you **a prot**ester
+5.288	a bo	in **a Bo**ld Display of Solidarity	-4.010	a crisi	Another lost life . We have **a crisi**s .
+5.196	abuse f	by police **abuse f**or decades	-3.930	a cris	deal with **a cris**is Why won't Romney
+5.088	an expr	it's **an expr**ession of pain... Baltimore	-3.667	a mur	don't support **a mur**derer and felon
+4.953	an hou	Just spent **an hour** weeping	-3.620	ahme	Liaqat incite hatred for **Ahme**dis in
+4.802	an effo	**an effo**rt to get food to victims	-3.496	a fun	**A fun**damental respect ... Taking lives
+4.782	a fag	i really **a fag** if I'm liking that video	-3.483	a viol	**A viol**ation of allegiance ... The betrayal

Autoridade

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+6.603	and al	during Sandy **and al**ways leadership	-6.462	arse	hey dumb **arse**...with such disrespect!
+6.398	all at	NO DACA **all at** the SAME TIME	-3.845	anal po	queer **anal po**unding
+5.864	a tin	grandfather who ... was **a tin**ner	-3.815	anal p	queer **anal p**ounding
+5.557	african	Pray for the **African**-American	-3.673	ars	**ars**onists justify destroying
+5.507	and ask	only person who can help us **and ask**	-3.407	a chick	Hate when **a chick** say
+5.348	aint bo	if you **aint bo**ut that Murder Game	-3.309	a hot	he isn't **a hot**shot any more
+5.314	are ne	Or you **are ne**ck deep with	-3.242	a dist	**a dist**raught grieving family
+5.314	a phil	stoicism ... **a phil**osophy of happiness	-3.185	amp th	**amp th**is one...served him for free
+5.302	ass fo	**ass fo**r fed taxpayer money	-2.997	and shu	police station **and shu**t down street
+5.149	are eno	mere words **are eno**ugh to offend you	-2.941	be piss	girl **be piss**ing me off all day

Pureza

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+7.583	a pub	of God's anger ... **a pub**lic toilet	-4.145	bad tod	to choke so **bad tod**ay
+5.834	and i d	**AND I d**on't condone violent protest	-4.025	and ren	**and ren**ounce pro-abortion stance
+5.823	and she	**and she** is funny, ultra passionate	-3.426	bad ton	drugs doesn't sound **bad ton**ight
+5.413	angles	it's all about the **angles**	-3.422	a lil	**a lil** juice and these bitches
+5.405	and tw	four bankruptcies **and tw**o divorces	-3.342	at my	honor law enforcement **at my** church
+5.223	a publ	of God's anger ... **a pub**lic toilet	-3.262	albi	Is that an **albi**no Mexican?
+5.042	any di	**any di**stance for all you love birds	-3.197	a gri	is **a gri**m, nasty, sadistic film
+4.798	an eu	not **an Eu**ropean... land of indigenous	-3.105	autis	who mocked an **autis**tic child
+4.742	arms	God is thy refuge ... everlasting **arms**	-3.017	ass ch	got **ass ch**eeks on my white tee
+4.495	bad wbu	not too **bad wbu**	-3.010	and gif	**and gif**t of your Son

Fonte: Alex Gwo Jen Lan, 2022

Tabela 24 – Palavras com maior influência no BRMoral em cada fundamento

Cuidado

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+14.905	aumenta	e **aumenta**r a qualidade de vida	-12.487	beb	e **beb**endo às nossas custas
+12.584	a indús	**indús**tria da carne um crime	-12.475	isso de	**isso de**manda um investimento
+11.734	flag	**flag**rante não deveriam nem ter direito	-12.475	investi	isso demanda um **investi**mento
+11.162	ilusó	noção de certeza de justiça é **ilusó**ria	-12.333	há vol	inocente depois, não **há vol**ta
+10.815	crimes	favorável ... a pessoas ...**crimes** hediondos	-11.551	confu	não vamos **confu**ndir as coisas
+10.485	a legal	**a legal**ização da morte ... é um risco	-11.551	confron	**confron**ta um dos princípios
+9.955	aumento	menos punitiva, ... gera um **aumento**	-10.877	laran	delações premiadas, **laran**jas
+9.955	aus	**aus**ência ou sucesso de recuperação	-10.848	confun	não vamos **confun**dir as coisas
+9.955	ausê	**ausê**ncia ou sucesso de recuperação	-10.848	confund	não vamos **confund**ir as coisas
+9.614	cumprim	recuperação ... **cumprim**ento da pena	-10.831	a pró	justo pagar com **a pró**pria vida

Justiça

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+1.634	o pad	Independentemente do ... **o pad**rão	-1.835	atu	deve promover ... **atu**almente
+1.624	aprese	**aprese**ntarem ... de forma burocrática	-1.681	aos seu	assegurar esse direito **aos seu**s
+1.612	inofen	é **inofen**sivo a sociedade	-1.584	a ótic	**a ótic**a restrita de um contrato
+1.528	exceção	com **exceção** das pessoas que	-1.534	a ótica	**a ótica** restrita de um contrato
+1.522	inofe	é **inofe**nsivo a sociedade	-1.534	e até	**e até** que me provem que o casamento
+1.467	ato pre	este **ato pre**judica na concepção	-1.523	a comun	sem consequências ... **a comun**idade
+1.467	ato pr	este **ato pr**ejudica na concepção	-1.503	e at	**e até** que me provem que o casamento
+1.419	na reg	Ajudaria **na reg**ulamentação	-1.440	anormai	não são **anormai**s ou doentes
+1.410	irmãos	ou **irmãos** já adultos que vivem	-1.391	de pai	aceitar uma família **de pai**s do
+1.373	ato pu	com um **ato pu**ramente burocrático	-1.355	judici	ser assegurada **judici**almente

Lealdade

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+8.730	da pe	grau psíquico e social **da pe**ssoa	-8.738	ap	-
+8.499	las co	pe**las co**nsequências que ele infringir	-7.758	a o cr	do que ... segurança contr**a o cr**ime.
+7.989	domicil	diminuir ... invasões **domicil**iares	-7.537	e pela	**e pela** forma como a sociedade
+7.558	além do	**Além do** mais, se fosse permitido	-7.286	cog	humano sequer **cog**itasse ferir ou matar
+7.446	dentro	to **dentro** de si	-7.036	discuss	traição do cônjuge, **discuss**ão em baladas
+7.167	e discu	**e discu**ssões cotidianas em fatalidades	-6.955	apa	pessoa **apa**rentemente qualificada
+7.142	e como	**e como** mulher, isso me preocupa	-6.928	a satis	se voltar contra ... não **a satis**faz
+7.086	impro	**impro**vável que um armamento	-6.835	e ouvi	no pouco que já li **e ouvi**
+7.084	bem ta	cidadão do **bem ta**mbém poderia	-6.804	ao mes	**ao mes**mo tempo ... em residências
+6.866	humanit	discernimento lógico e **humanit**ário	-6.171	iria r	e não **iria r**esolver o problema

Autoridade

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+14.079	certa s	opinião **certa s**obre ... com os mais velhos	-13.675	mais fa	**mais fa**vorável até para
+12.534	consigo	Não **consigo** perceber a diferença	-12.754	apoiand	não é uma boa medida me **apoiand**o
+11.701	aprese	**aprese**ntada por diversos atores	-12.474	crime a	"recrutados"para o **crime a**inda mais
+11.701	certas	já pode tomar **certas** decisões	-12.312	de coe	é uma questão **de coe**rência
+11.081	a capa	**a capa**cidade do adolescente	-11.524	lado d	do **lado d**e criminosos de verdade
+11.021	isso a	Por **isso a** redução de dois anos	-11.456	da su	ação para o restante **da su**a vida
+10.810	a possi	não descarta **a possi**bilidade de recuperação	-10.698	assass	pai de família, que é **assass**inado por
+10.394	algo me	estado secundário, ... como **algo me**lhor	-10.694	estão	do Estado. Os presídios já **estão**
+10.362	como pe	**como pe**ões do crime	-10.442	e fer	criminalidade **e fer**e os direitos do outro
+9.797	ciclo e	crescido mais rápido, o **ciclo e**stá	-10.272	ao a	Estado levarão **ao a**umento da violência

Pureza

Peso(+)	N-grama	Exemplo	Peso(-)	N-grama	Exemplo
+14.256	iguai	Todos são **iguai**s perante a lei	-12.809	e ci	condições **e ci**rcunstâncias, doenças
+12.345	amiga	uma **amiga** que abortou e se arrepende	-12.189	elas da	se **elas da**riam essa chance a essas mães
+12.208	a nas	até que a criança viesse **a nas**cer	-11.007	da pu	O homem fica livre **da pu**nição
+11.568	já são	a proteção a mulher ... **já são** legalizados	-10.289	lhe a	sem dar-**lhe a** chance de uma vida plena
+10.799	estudi	e epidemias, **estudi**osos do tema	-9.805	de proc	legalização, o número **de pr**ocedimentos
+10.602	esta	**esta** por sua vez tem um valor único para Deus	-9.414	lhe	consome seus nutrientes, **lhe** causa dores
+10.114	fazer u	direito de **fazer u**ma laqueadura quando	-9.342	banais	favorável em casos **banais**
+9.792	deve to	somente ela, **deve to**mar as decisões	-9.180	lhe a c	retirar a criança sem dar-**lhe a c**hance
+9.568	já li n	**já li n**otícias de que a incidência	-9.083	em se	psicológicas quanto **em se**u corpo
+9.379	as nece	**as nece**ssidades básicas da saúde	-9.046	a pess	que **a pess**oa sofreu um abuso

Fonte: Alex Gwo Jen Lan, 2022

5.4.2 Aprendizado de fundamentos morais pessoais no córpus BRmoral

Apesar do desempenho razoável do BERT/ELMo, é digno de nota que *reglog.char* apresentou resultados consideravelmente superiores para as tarefas de CFMI e CFMP. Assim, como forma de investigar esta questão em mais detalhes, foi conduzida uma análise manual de exemplos de classificação do córpus BRmoral em que *reglog.char* e BERT/ELMo apresentam resultados divergentes do esperado. Um exemplo deste tipo - uma opinião sobre o tópico "pena de morte"extraído do córpus BRMoral - é ilustrado a seguir:

> *Exemplo: Tenho pouca opinião ou conhecimento sobre o tema. A depender da gravidade do crime cometido, pode ser uma consequência eficaz para minimização. Por outro lado, alguns Estados desenvolvidos não tem lançado mão desse meio de punição e ainda assim tem índices pouco expressivos de determinados tipos de crimes. São feitos investimentos na área social e da educação, em contrapartida.*

No exemplo mencionado, *reglog.char* classificou corretamente com o rótulo "*high*"para Cuidado, mas BERT/ELMo erraram ao classificar com o rótulo "*low*"para o mesmo fundamento. Uma possível explicação para isso pode estar associada à estrutura das opiniões no córpus, que apresenta alternância de argumentos contra e a favor do assunto em discussão como neste exemplo. Em situações de alta complexidade como esta, é possível que o córpus não apresente um número suficiente de instâncias para a modelagem do problema como uma tarefa de classificação de sequências (ao estilo feito pelos modelos neurais), o que por outro lado representaria uma vantagem para abordagens mais simples baseadas na contagem de *tokens* (ao estilo dos modelos *reglog.char* e outros). Consideramos entretanto que uma conclusão mais definitiva a esse respeito exija novos estudos.

6 Conclusões

A identificação de valores morais em discursos humanos tem um papel importante na compreensão de conflitos sociais, além de comportamentos e posições ideológicas frente a uma diversidade de assuntos. E como apresentado na revisão bibliográfica, este tópico é de grande interesse para várias aplicações na ciência política, psicologia e computação. Assim, em vista da importância deste tipo de tarefa, formulou-se uma proposta de pesquisa em nível de mestrado acadêmico para o desenvolvimento de abordagens computacionais integrando a Teoria dos Fundamentos Morais (TFM). No presente estudo, investigou-se a possibilidade de abordar a tarefa de classificação de fundamentos morais a partir de modelos pré-treinados de *embeddings* contextuais.

Os experimentos realizados sugerem resultados que contribuem para o avanço do estado-da-arte em classificação de fundamentos morais. Para as tarefas de *classificação de fundamentos morais impessoais*, ou CFMI, demonstrou-se a eficácia do uso de *embeddings* contextuais e também da regressão logística com n-gramas de caracteres. Para as tarefas de *classificação de fundamentos morais pessoais*, ou CFMP, embora o modelo de *embeddings* contextuais não tenha alcançado os melhores resultados, ainda foi possível desenvolver modelos para este tipo de tarefa, que são inéditos para a literatura, com regressão logística com n-gramas de caracteres. As principais contribuições deste estudo são listadas abaixo.

- Modelos de CFMI baseados no pré-treinamento de *embeddings* contextuais e regressão logística com n-gramas a nível de caracteres no idioma em inglês, validados no domínio de *tweets*.
- Modelos de CFMP baseados em regressão logística com n-gramas a nível de caracteres em idioma português brasileiro, validados no domínio de textos de opinião.

Em vista das limitações encontradas nos experimentos de CFMP, surgem oportunidades de melhoria, que também são válidas para a CFMI. De forma mais específica, sugere-se como trabalho futuro a exploração da tarefa a partir de estratégias que envolvam *stacking* de classificadores (WOLPERT, 1992), ou a combinação de técnicas e outras características para a classificação (HULPUS *et al.*, 2020; ROY; GOLDWASSER, 2021), além de novos conjuntos de dados rotulados com informação moral em desenvolvimento (JOHNSON; JIN; GOLDWASSER, 2017) durante a elaboração desta pesquisa.

Referências[1]

AKBIK, A.; BERGMANN, T.; BLYTHE, D.; RASUL, K.; SCHWETER, S.; VOLLGRAF, R. FLAIR: an easy-to-use framework for state-of-the-art NLP. In: AMMAR, W.; LOUIS, A.; MOSTAFAZADEH, N. (Ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. p. 54–59. Citado na página 60.

AMIN, A. B.; BEDNARCZYK, R. A.; RAY, C. E.; MELCHIORI, K. J.; GRAHAM, J.; HUNTSINGER, J. R.; OMER, S. B. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, Nature Publishing Group, v. 1, n. 12, p. 873–880, 2017. Citado na página 12.

ARAQUE, O.; GATTI, L.; KALIMERI, K. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, Elsevier, v. 191, p. 105184, 2020. Citado 7 vezes nas páginas 19, 49, 56, 61, 62, 64 e 70.

BENGIO, Y.; SIMARD, P. Y.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, IEEE, v. 5, n. 2, p. 157–166, 1994. Citado na página 27.

CARVALHO, F.; OKUNO, H. Y.; BARONI, L.; GUEDES, G. P. A brazilian portuguese moral foundations dictionary for fake news classification. In: *39th International Conference of the Chilean Computer Science Society*. Coquimbo, Chile: IEEE, 2020. p. 1–5. Citado na página 19.

CASTRO, P. V. Q. de. *Deep Learning for Named Entity Recognition in Legal Domain*. Dissertação (Mestrado) — Universidade Federal de Goias, Goiania, Brazil, 01 2019. Citado 3 vezes nas páginas 33, 60 e 61.

CHO, K.; MERRIENBOER, B. van; GÜLÇEHRE, Ç.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1724–1734. Citado na página 27.

CLIFFORD, S.; JERIT, J. How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, Cambridge University Press New York, USA, v. 75, n. 3, p. 659–671, 2013. Citado na página 37.

CURRY, O. S.; MULLINS, D. A.; WHITEHOUSE, H. Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, The University of Chicago Press Chicago, IL, v. 60, n. 1, p. 47–69, 2019. Citado na página 12.

DAVIDSON, T.; WARMSLEY, D.; MACY, M. W.; WEBER, I. Automated hate speech detection and the problem of offensive language. In: *Proceedings of the Eleventh International Conference on Web and Social Media*. Montréal, Québec, Canada: AAAI Press, 2017. p. 512–515. Citado na página 21.

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

DEERWESTER, S. C.; DUMAIS, S. T.; LANDAUER, T. K.; FURNAS, G. W.; HARSHMAN, R. A. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990. Citado na página 39.

DEHGHANI, M. Purity homophily in social networks - invited talk. In: BALAHUR, A.; GOOT, E. V. der; VOSSEN, P.; MONTOYO, A. (Ed.). *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California, USA: The Association for Computer Linguistics, 2016. p. 16. Citado 4 vezes nas páginas 29, 40, 41 e 56.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. p. 4171–4186. Citado 8 vezes nas páginas 14, 33, 34, 35, 36, 59, 60 e 61.

DINKOV, Y.; KOYCHEV, I.; NAKOV, P. Detecting toxicity in news articles: application to bulgarian. In: MITKOV, R.; ANGELOVA, G. (Ed.). *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria: INCOMA Ltd., 2019. p. 247–258. Citado na página 53.

ELMAN, J. L. Finding structure in time. *Cognitive Science*, Wiley Online Library, v. 14, n. 2, p. 179–211, 1990. Citado na página 25.

FULGONI, D.; CARPENTER, J.; UNGAR, L. H.; PREOTIUC-PIETRO, D. An empirical exploration of moral foundations theory in partisan news sources. In: CALZOLARI, N.; CHOUKRI, K.; DECLERCK, T.; GOGGI, S.; GROBELNIK, M.; MAEGAARD, B.; MARIANI, J.; MAZO, H.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 3730–3736. Citado 2 vezes nas páginas 12 e 38.

GANITKEVITCH, J.; DURME, B. V.; CALLISON-BURCH, C. PPDB: the paraphrase database. In: VANDERWENDE, L.; III, H. D.; KIRCHHOFF, K. (Ed.). *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*. Atlanta, Georgia, USA: The Association for Computational Linguistics, 2013. p. 758–764. Citado na página 47.

GARTEN, J.; BOGHRATI, R.; HOOVER, J.; JOHNSON, K. M.; DEHGHANI, M. Morality between the lines: Detecting moral sentiment in text. In: *Proceedings IJCAI 2016 workshop on Computational Modeling of Attitudes*. New York, USA: Association for the Advancement of Artificial Intelligence, 2016. Citado na página 63.

GARTEN, J.; HOOVER, J.; JOHNSON, K. M.; BOGHRATI, R.; ISKIWITCH, C.; DEHGHANI, M. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, Springer, v. 50, n. 1, p. 344–361, 2018. Citado 4 vezes nas páginas 13, 14, 42 e 56.

GOODFELLOW, I. J.; BENGIO, Y.; COURVILLE, A. C. *Deep Learning*. Cambridge: MIT Press, 2016. (Adaptive computation and machine learning). ISBN 978-0-262-03561-3. Citado na página 24.

GRAHAM, J.; HAIDT, J.; KOLEVA, S.; MOTYL, M.; IYER, R.; WOJCIK, S. P.; DITTO, P. H. Moral foundations theory: The pragmatic validity of moral pluralism. In: DEVINE, A. P. P. (Ed.). *Advances in experimental social psychology*. [S.l.]: Academic Press, 2013. v. 47, p. 55–130. Citado 4 vezes nas páginas 12, 16, 17 e 18.

GRAHAM, J.; HAIDT, J.; NOSEK, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, American Psychological Association, v. 96, n. 5, p. 1029–1046, 2009. Citado 5 vezes nas páginas 12, 18, 19, 37 e 62.

GRAHAM, J.; NOSEK, B. A.; HAIDT, J.; IYER, R.; KOLEVA, S.; DITTO, P. H. Mapping the moral domain. *Journal of Personality and Social Psychology*, American Psychological Association, v. 101, n. 2, p. 366–385, 2011. Citado 2 vezes nas páginas 18 e 22.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 27.

HOOVER, J.; JOHNSON, K.; BOGHRATI, R.; GRAHAM, J.; DEHGHANI, M.; DONNELLAN, M. B. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, University of California Press, v. 4, n. 1, p. 1–18, 2018. Citado na página 12.

HOOVER, J.; PORTILLO-WIGHTMAN, G.; YEH, L.; HAVALDAR, S.; DAVANI, A. M.; LIN, Y.; KENNEDY, B.; ATARI, M.; KAMEL, Z.; MENDLEN, M.; MORENO, G.; PARK, C.; CHANG, T. E.; CHIN, J.; LEONG, C.; LEUNG, J. Y.; MIRINJIAN, A.; DEHGHANI, M. Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, SAGE Publications Sage CA: Los Angeles, CA, v. 11, n. 8, p. 1057–1071, 2020. Citado 5 vezes nas páginas 12, 15, 20, 21 e 61.

HOPP, F. R.; FISHER, J. T.; CORNELL, D.; HUSKEY, R.; WEBER, R. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, Springer, v. 53, n. 1, p. 232–246, 2020. Citado na página 19.

HULPUS, I.; KOBBE, J.; STUCKENSCHMIDT, H.; HIRST, G. Knowledge graphs meet moral values. In: GUREVYCH, I.; APIDIANAKI, M.; FARUQUI, M. (Ed.). *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. Barcelona, Spain: Association for Computational Linguistics, 2020. p. 71–80. Citado 2 vezes nas páginas 54 e 80.

JOHNSON, K.; GOLDWASSER, D. Classification of moral foundations in microblog political discourse. In: GUREVYCH, I.; MIYAO, Y. (Ed.). *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 720–730. Citado 3 vezes nas páginas 46, 48 e 56.

JOHNSON, K.; GOLDWASSER, D. Modeling behavioral aspects of social media discourse for moral classification. In: VOLKOVA, S.; JURGENS, D.; HOVY, D.; BAMMAN, D.; TSUR, O. (Ed.). *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. p. 100–109. Citado 2 vezes nas páginas 48 e 56.

JOHNSON, K.; JIN, D.; GOLDWASSER, D. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter. In: BARZILAY, R.; KAN, M. (Ed.). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 741–752. Citado 4 vezes nas páginas 46, 48, 52 e 80.

JOSHI, A.; KARIMI, S.; SPARKS, R.; PARIS, C.; MACINTYRE, C. R. A comparison of word-based and context-based representations for classification problems in health informatics. In: DEMNER-FUSHMAN, D.; COHEN, K. B.; ANANIADOU, S.; TSUJII, J. (Ed.). *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, 2019. p. 135–141. Citado 2 vezes nas páginas 14 e 33.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing (draft). october 2019. *URL https://web. stanford. edu/˜ jurafsky/slp3*, 2019. Citado 3 vezes nas páginas 23, 24 e 28.

KALIMERI, K.; BEIRÓ, M. G.; DELFINO, M.; RALEIGH, R.; CATTUTO, C. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, Elsevier, v. 92, p. 428–445, 2019. Citado na página 12.

KAUR, R.; SASAHARA, K. Quantifying moral foundations from various topics on twitter conversations. In: JOSHI, J.; KARYPIS, G.; LIU, L.; HU, X.; AK, R.; XIA, Y.; XU, W.; SATO, A.; RACHURI, S.; UNGAR, L. H.; YU, P. S.; GOVINDARAJU, R.; SUZUMURA, T. (Ed.). *2016 IEEE International Conference on Big Data*. Washington DC, USA: IEEE Computer Society, 2016. p. 2505–2512. Citado 3 vezes nas páginas 29, 40 e 56.

KOLEVA, S. P.; GRAHAM, J.; IYER, R.; DITTO, P. H.; HAIDT, J. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, Elsevier, v. 46, n. 2, p. 184–194, 2012. Citado 2 vezes nas páginas 12 e 18.

LE, Q. V.; MIKOLOV, T. Distributed representations of sentences and documents. In: *Proceedings of the 31th International Conference on Machine Learning*. Beijing, China: JMLR.org, 2014. (JMLR Workshop and Conference Proceedings, v. 32), p. 1188–1196. Citado na página 41.

LIN, Y.; HOOVER, J.; PORTILLO-WIGHTMAN, G.; PARK, C.; DEHGHANI, M.; JI, H. Acquiring background knowledge to improve moral value prediction. In: BRANDES, U.; REDDY, C.; TAGARELLI, A. (Ed.). *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining*. Barcelona, Spain: IEEE Computer Society, 2018. p. 552–559. Citado 4 vezes nas páginas 13, 14, 45 e 56.

LOW, M.; WUI, M. G. L. Moral foundations and attitudes towards the poor. *Current Psychology*, Springer, v. 35, n. 4, p. 650–656, 2016. Citado na página 12.

MATSUO, A.; SASAHARA, K.; TAGUCHI, Y.; KARASAWA, M. Development and validation of the japanese moral foundations dictionary. *PloS one*, Public Library of Science, v. 14, n. 3, p. e0213343, 2019. Citado na página 19.

MCCLELLAND, J. L.; RUMELHART, D. E.; GROUP, P. R. *et al.* Parallel distributed processing. *Explorations in the Microstructure of Cognition*, MIT Press Cambridge, Ma, v. 2, p. 216–271, 1986. Citado na página 25.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, Springer, v. 12, n. 2, p. 153–157, 1947. Citado 2 vezes nas páginas 68 e 69.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). *1st International Conference on Learning Representations*. Scottsdale, Arizona, USA: [s.n.], 2013. Citado 3 vezes nas páginas 14, 30 e 43.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: BURGES, C. J. C.; BOTTOU, L.; GHAHRAMANI, Z.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. Lake Tahoe, Nevada, USA: [s.n.], 2013. p. 3111–3119. Citado 2 vezes nas páginas 31 e 32.

MIKOLOV, T.; YIH, W.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: VANDERWENDE, L.; III, H. D.; KIRCHHOFF, K. (Ed.). *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*. Atlanta, Georgia, USA: The Association for Computational Linguistics, 2013. p. 746–751. Citado 2 vezes nas páginas 30 e 32.

MOOIJMAN, M.; HOOVER, J.; LIN, Y.; JI, H.; DEHGHANI, M. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, Nature Publishing Group, v. 2, n. 6, p. 389–396, 2018. Citado na página 53.

NIELSEN, M. A. *Neural networks and deep learning*. [S.l.]: Determination press San Francisco, CA, 2015. v. 2018. Citado na página 25.

NOKHIZ, P.; LI, F. Understanding rating behavior based on moral foundations: The case of yelp reviews. In: NIE, J.; OBRADOVIC, Z.; SUZUMURA, T.; GHOSH, R.; NAMBIAR, R.; WANG, C.; ZANG, H.; BAEZA-YATES, R.; HU, X.; KEPNER, J.; CUZZOCREA, A.; TANG, J.; TOYODA, M. (Ed.). *2017 IEEE International Conference on Big Data*. Boston, MA, USA: IEEE, 2017. p. 3938–3945. Citado 3 vezes nas páginas 14, 41 e 56.

PACHECO, M. L.; GOLDWASSER, D. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 9, p. 100–119, 2021. Citado na página 51.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. *The PageRank Citation Ranking: Bringing Order to the Web*. [S.l.], 1999. Previous number = SIDL-WP-1999-0120. Disponível em: ⟨http://ilpubs.stanford.edu:8090/422/⟩. Citado na página 54.

PAVAN, M. C.; SANTOS, V. G. dos; LAN, A. G. J.; MARTINS, J. ao T.; SANTOS, W. R. dos; DEUTSCH, C.; COSTA, P. B. da; HSIEH, F. C.; PARABONI, I. Morality classification in natural language text. *IEEE transactions on Affective Computing*, IEEE, 2020. Citado 5 vezes nas páginas 15, 20, 21, 23 e 57.

PECAR, S.; SIMKO, M.; BIELIKOVÁ, M. Improving sentiment classification in slovak language. In: ERJAVEC, T.; MARCINCZUK, M.; NAKOV, P.; PISKORSKI, J.; PIVOVAROVA, L.; SNAJDER, J.; STEINBERGER, J.; YANGARBER, R. (Ed.). *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Florence,

Italy: Association for Computational Linguistics, 2019. p. 114–119. Citado 2 vezes nas páginas 14 e 33.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, JMLR.org, v. 12, p. 2825–2830, 2011. Citado na página 63.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Citado na página 43.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana USA: Association for Computational Linguistics, 2018. p. 2227–2237. Citado 5 vezes nas páginas 14, 33, 59, 60 e 61.

RANGEL, F.; ROSSO, P. Overview of the 7th Author profiling task at PAN 2019: bots and gender profiling. In: CAPPELLATO, L.; FERRO, N.; LOSADA, D.; MÜLLER, H. (Ed.). *CLEF 2019 Labs and Workshops, Notebook Papers*. Lugano, Switzerland: CEUR-WS.org, 2019. p. 36. Citado 2 vezes nas páginas 13 e 57.

RANGEL, F.; ROSSO, P.; ZAGHOUANI, W.; CHARFI, A. Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, Cambridge University Press, v. 26, n. 6, p. 641–661, 2020. Citado na página 13.

REZAPOUR, R.; SHAH, S. H.; DIESNER, J. Enhancing the measurement of social effects by capturing morality. In: BALAHUR, A.; KLINGER, R.; HOSTE, V.; STRAPPARAVA, C.; CLERCQ, O. D. (Ed.). *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. p. 35–45. Citado na página 19.

ROBERTSON, F.; LAGUS, J.; KAJAVA, K. A COVID-19 news coverage mood map of europe. In: TOIVONEN, H.; BOGGIA, M. (Ed.). *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Online: Association for Computational Linguistics, 2021. p. 110–115. Citado na página 54.

ROY, S.; GOLDWASSER, D. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In: KU, L.; LI, C. (Ed.). *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, 2021. p. 1–13. Citado 3 vezes nas páginas 51, 56 e 80.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning internal representations by error propagation*. [S.l.], 1985. Citado na página 25.

SAGI, E.; DEHGHANI, M. Measuring moral rhetoric in text. *Social science computer review*, Sage Publications Sage CA: Los Angeles, CA, v. 32, n. 2, p. 132–144, 2014. Citado 4 vezes nas páginas 29, 39, 41 e 56.

SANTOS, W. R. dos; RAMOS, R. M. S.; PARABONI, I. Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, Taylor & Francis, v. 25, n. 4, p. 268–287, 2019. Citado na página 13.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, IEEE, v. 45, n. 11, p. 2673–2681, 1997. Citado na página 26.

SILVINO, A. M. D.; PILATI, R.; KELLER, V. N.; SILVA, E. P.; FREITAS, A. F. de P.; SILVA, J. N.; LIMA, M. F. Adaptac ao do questionário dos fundamentos morais para o português. *Psico-USF*, SciELO Brasil, v. 21, n. 3, p. 487–495, 2016. Citado 3 vezes nas páginas 13, 18 e 57.

SMITH, H. M. Interpreting qualitative data: methods for analyzing talk, text and interaction 3rd edition. *Sociological Research Online*, SAGE Publications Sage UK: London, England, v. 11, n. 4, p. 100–101, 2006. Citado na página 37.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation. In: SATTAR, A.; KANG, B. (Ed.). *AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence*. Hobart, Australia: Springer, 2006. (Lecture Notes in Computer Science, v. 4304), p. 1015–1021. Citado na página 68.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *9th Brazilian Conference on Intelligent Systems (BRACIS)*. Rio Grande, Brasil: Springer, 2020. v. 12319, p. 403–417. Citado 2 vezes nas páginas 33 e 61.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE, N. D.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*. Montreal, Quebec, Canada: [s.n.], 2014. p. 3104–3112. Citado 2 vezes nas páginas 27 e 28.

SYLWESTER, K.; PURVER, M. Twitter language use reflects psychological differences between democrats and republicans. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 9, p. e0137422, 2015. Citado 2 vezes nas páginas 12 e 38.

TAKAHASHI, T.; TAHARA, T.; NAGATANI, K.; MIURA, Y.; TANIGUCHI, T.; OHKUMA, T. Text and image synergy with feature cross technique for gender identification: notebook for PAN at CLEF 2018. In: CAPPELLATO, L.; FERRO, N.; NIE, J.; SOULIER, L. (Ed.). *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. Avignon, France: CEUR-WS.org, 2018. (CEUR Workshop Proceedings, v. 2125). Citado na página 13.

TAKIKAWA, H.; SAKAMOTO, T. Moral foundations of political discourse: Comparative analysis of the speech records of the us congress and the japanese diet. *arXiv preprint arXiv:1704.06903*, 2017. Citado na página 53.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010. Citado na página 37.

TEERNSTRA, L.; PUTTEN, P. van der; NOORDEGRAAF-EELENS, L.; VERBEEK, F. J. The morality machine: tracking moral values in tweets. In: BOSTRÖM, H.; KNOBBE, A. J.; SOARES, C.; PAPAPETROU, P. (Ed.). *Advances in Intelligent Data Analysis XV - 15th International Symposium*. Stockholm, Sweden: Springer, 2016. (Lecture Notes in Computer Science, v. 9897), p. 26–37. Citado 6 vezes nas páginas 13, 14, 23, 29, 44 e 56.

VOLKOVA, S.; SHAFFER, K.; JANG, J. Y.; HODAS, N. O. Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter. In: BARZILAY, R.; KAN, M. (Ed.). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 647–653. Citado na página 53.

WANG, H.; LIU, C.; YU, D. (construction of a chinese moral dictionary for artificial intelligence ethical computing). In: SUN, M.; LI, S.; ZHANG, Y.; LIU, Y. (Ed.). *Proceedings of the 19th Chinese National Conference on Computational Linguistics*. Haikou, China: Chinese Information Processing Society of China, 2020. p. 539–549. Citado na página 19.

WANG, X.; SHI, W.; KIM, R.; OH, Y.; YANG, S.; ZHANG, J.; YU, Z. Persuasion for good: towards a personalized persuasive dialogue system for social good. In: KORHONEN, A.; TRAUM, D. R.; MÀRQUEZ, L. (Ed.). *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 5635–5649. Citado na página 53.

WINTERICH, K. P.; ZHANG, Y.; MITTAL, V. How political identity and charity positioning increase donations: Insights from moral foundations theory. *International Journal of Research in Marketing*, Elsevier, v. 29, n. 4, p. 346–354, 2012. Citado na página 12.

WOLPERT, D. H. Stacked generalization. *Neural networks*, v. 5, n. 2, p. 241–259, 1992. Citado na página 80.

WOLSKO, C.; ARICEAGA, H.; SEIDEN, J. Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, Elsevier, v. 65, p. 7–19, 2016. Citado na página 12.

XIE, J. Y.; JUNIOR, R. F. P.; HIRST, G.; XU, Y. Text-based inference of moral sentiment change. In: INUI, K.; JIANG, J.; NG, V.; WAN, X. (Ed.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 4653–4662. Citado na página 54.

ZAMPIERI, M.; MALMASI, S.; NAKOV, P.; ROSENTHAL, S.; FARRA, N.; KUMAR, R. Semeval-2019 task 6: identifying and categorizing offensive language in social media (offenseval). In: MAY, J.; SHUTOVA, E.; HERBELOT, A.; ZHU, X.; APIDIANAKI, M.; MOHAMMAD, S. M. (Ed.). *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. p. 75–86. Citado 2 vezes nas páginas 14 e 33.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1253, 2018. Citado na página 57.