UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE EDUCAÇÃO ONLINE
ESPECIALIZAÇÃO EM BIG DATA, DATA SCIENCE E DATA ANALYTICS

Carla Machado Ferreira

IDENTIFICAÇÃO DE DISCURSO DE ÓDIO ATRAVÉS DE ALGORITMOS DE MACHINE LEARNING

Porto Alegre
2019

UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE EDUCAÇÃO ONLINE
ESPECIALIZAÇÃO EM BIG DATA, DATA SCIENCE E DATA ANALYTICS

Carla Machado Ferreira

IDENTIFICAÇÃO DE DISCURSO DE ÓDIO ATRAVÉS DE ALGORITMOS DE MACHINE LEARNING

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Especialista em Big Data, Data Science e Data Analytics, pelo curso de Pós-Graduação Lato Sensu em Big Data, Data Science e Data Analytics da Universidade do Vale do Rio dos Sinos – UNISINOS.

Orientador: Prof. Dra. Patricia Milard Jacques

Porto Alegre
2019

Identificação de Discurso de Ódio através de algoritmos de machine learning

Carla Machado Ferreira

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Nilo Peçanha, 1600, Boa Vista, Porto Alegre, RS – Brasil

carlamf@gmail.com

Abstract. *The growth of social networks brought as negative factor the spread of hate speech, which impacts society negatively. Due to the large volume of shared comments, manual classification becomes infeasible and the use of computational techniques can assist in the automatic identification of offensive opinions in Portuguese. This article demonstrates the analysis of Supervised Learning methods for rating hate speech in Portuguese. More specifically, techniques already demonstrated in English language research with good results will be employed. The results for F1-Score showed viable the application of techniques such as Mutilayer Perceptrons (80 %) and Random Forest (77 %).*

Resumo. *O crescimento das redes sociais trouxe como fator negativo a propagação de discurso de ódio, que impacta a sociedade negativamente. Devido ao grande volume de comentários compartilhadas, a classificação manual se torna inviável e o uso de técnicas computacionais podem auxiliar na identificação automática de opiniões ofensivas em português. Este artigo demonstra a análise de métodos de Aprendizado Supervisionado para classificação de discurso de ódio de comentários em português. Mais especificamente, serão empregadas técnicas já demonstradas em pesquisas na língua inglesa com bons resultados. Os resultados para F1-Score mostraram viáveis a aplicação de técnicas como Mutilayer Perceptrons (80%) e Random Forest (77%).*

1. Introdução

O dinamismo e alcance global proporcionado pela Internet aliado ao rápido crescimento das redes sociais promoveram uma nova configuração de comunicação online com uma maior interação entre usuários, compartilhamento de conteúdo e um meio de expressar opiniões de maneira espontânea. No entanto, o anonimato e mobilidade permitidos em plataformas digitais [Malmasi and Zampieri 2017], associados à crença de impunidade, propiciam um comportamento agressivo e uma ferramenta para propagação do discurso de ódio.

Números sugerem um crescimento de publicações contendo mensagens de intolerância a indivíduos ou determinados grupos na Internet Brasileira. Nos últimos treze anos, a Central Nacional de Denúncias de Crimes Cibernéticos recebeu cerca de quatro milhões de denúncias [Safernet 2019] , metade relacionados aos crimes de ódio como racismo, neonazismo, homofobia e intolerância religiosa. O discurso de ódio não afeta

apenas em ambientes virtuais, já que ataques realizados online têm grande probabilidade de estender no mundo real e trazer risco a sociedade. [Rothenburg and Stroppa 2015].

Ao expressar opiniões em redes sociais é possível expressar intolerância de forma pública e de certa forma validar o discurso de ódio com outros indivíduos que compartilham deste mesmo pensamento. Este tipo de comportamento pode criar uma base para a violência perigosa para a convivência em sociedade, uma vez que pode incentivar atos hostis ou violentos como, por exemplo, o assassinato do capoeirista Moa do Katendê por critivar um cadidato as eleições de 2018 [G1 2018], a ameaça em redes sociais que se concretizou em um ato de estupro contra uma estudante em Fortaleza [Nogueira 2018] ou mesmo ataque ao centro cultural árabe Al Janiah. [Martins 2019].

Devido ao grande volume de dados gerados, a identificação manual deste conteúdo contendo insultos ofensivos se torna uma tarefa impraticável [Pang et al. 2008]. Com o aumento da capacidade de processamento de grandes volumes de dados e para auxiliar na resolução do problema, tarefas de Processamento de Linguagem Natural (PLN) combinadas com técnicas de Aprendizado de Máquina objetivam entendimento semântico e classificação de textos de forma automática. [Abbasi et al. 2008].

O entendimento de emoções contidos em textos de comentários *online* impulsiona o desenvolvimento de mecanismos eficientes que contribuam na detecção da polaridade (positivo ou negativo) contido em cada sentença de publicações envolvendo discurso de ódio. Nos últimos anos, houve um crescimento de pesquisas que objetivam propor modelos computacionais que auxiliem na identificação do discurso de ódio. [Fortuna and Nunes 2018]. No entanto, a grande maioria destes estudos para detecção de discurso de ódio são em língua inglesa [Burnap and Williams 2015, Davidson et al. 2017, Badjatiya et al. 2017] e ainda há uma carência de pesquisas voltadas à língua portuguesa.

O único estudo proposto, até o momento, para identificação automática de discurso de ódio em comentários realizados no idioma português, foi realizado por Pelle e Moreira [2017]. No entanto, os experimentos foram conduzidos com os classificadores *Support Vector Machine* (SVN) [Han et al. 2011] e *Naive Bayes* (NB) [Russell and Norvig 2016]. Em razão da lacuna deixada no trabalho de Pelle e Moreira [2017], a análise da viabilidade de aplicação de outros métodos de classificação para identificação de ódio e que já foram utilizados com melhores resultados no idioma inglês possa a vir a contribuir na identificação de discurso de ódio em português.

Este artigo propõe a análise de técnicas de Aprendizado de Máquina que foram aplicados com resultados aceitáveis no idioma inglês e que até o presente momento não foram aplicados em trabalhos no idioma português. Devido à natureza experimental, foi adotado como método de pesquisa *Desing Science Search* para condução do presente trabalho. Esta forma de pesquisa objetiva compreender, explicar e identificar hipóteses associadas à pesquisa e em apoio à Tecnologia da Informação [Dresch et al. 2015]. Para tal finalidade, a pesquisa foi dividida em duas etapas: revisão bibliográfica e realização de experimento para que sejam compreendidos, avaliados e discutidos a aplicabilidade de técnicas adequadas para a solução do problema discutido.

Para a realização dos experimentos, foi utilizado um conjunto de dados contendo comentários da web em português que foram anotados manualmente como discurso de

ódio no trabalho conduzido por Pelle e Moreira [2017][1]. Posteriormente, foram explorados a aplicação dos métodos de classificação Regressão Logística, Árvore de Decisão, *Random Forest*, *Multilayer Perceptrons*, *Convolutional Neural Network* e *Long Short-Term Memory Network*, além das técnicas de classificação *Support Vector Machine* e *Naive Bayes*, também utilizados no trabalho de referência.

A experimentação desses outros métodos citados poderá vir a contribuir na identificação de discurso de ódio com uma maior precisão, pois já foram utilizados anteriormente com êxitos em experimentos para a identificação de comentários odiosos no idioma inglês [Wulczyn et al. 2017, Badjatiya et al. 2017, Waseem and Hovy 2016]. Dessa forma, podem ser uma alternativa aos modelos supostamente apresentados com eficiência por Pelle e Moreira [2017].

Este artigo está organizado conforme descrito a seguir: Na seção 1 são apresentadas o presente trabalho e a forma como foi organizado. A seção 2 demonstra a conceituação que auxiliará no entendimento do experimento realizado. Na seção 3 são apresentados trabalhos relacionados. A seção 4 apresenta a metodologia aplicada para o desenvolvimento dos experimentos. E, por fim, nas seções 5 e 6 são, respectivamente, apresentados e discutidos os resultados obtidos para que desta forma sejam realizadas as ponderações finais para a conclusão do trabalho.

2. Fundamentos Teóricos

Nesta seção é apresentada a fundamentação teórica que auxiliará no entendimento do problema de identificação de discurso de ódio em comentários textuais. Para este objetivo, são demonstrados os conceitos para o problema de discurso de ódio, os conceitos de descoberta do conhecimento, análise de sentimentos e mineração de opiniões, processamento de linguagem natural. Também é apresentado técnicas de aprendizado de máquina e, por fim, é realizada uma descrição das métricas de avaliação utilizadas para validar os resultados deste trabalho.

2.1. Discurso de Ódio

A velocidade e o alcance da Internet trouxeram um caráter transformador para a sociedade e um espaço propício para compartilhar opiniões e se expressar livremente. Esta mudança também promoveu a propagação de discurso de ódio em meio digital (*online*) de forma anônima, dificultando a identificação da autoria das mensagens de insulto e menos sujeito a controle do que na vida real (*off-line*). [Georgescu 2014].

O discurso de ódio, também denominado *hate speech*, é uma forma de abuso do direito de liberdade de expressão caracterizado pela exposição de mensagem de caráter discriminatório orientado a um indivíduo ou grupo com o intuito de ofender, intimidar e desvalorizar a vítima em virtude de sua raça, cor, nacionalidade, religião, orientação sexual ou gênero. [Horbach 2012].

De forma uma geral, o direito constituinte ou o direito internacional nem permite ou proíbe o discurso de ódio de forma consistente [Brugger 2007]. O entendimento do limite do direito de liberdade de expressão abre um questionamento da legitimidade da intervenção do Estado em invalidar discursos discriminatórios e ultrapassar direitos individuais. [Sarmento 2006].

[1]Disponível em https://github.com/rogersdepelle/OffComBR

Um Estado é composto por princípios morais e políticos que são adotados ao longo de sua história e com interpretações distintas da liberdade de expressão. Estados liberais tendem a valorizar a liberdade de expressão, mesmo com perda ao indivíduo. Por outro lado, Estados socialistas são propensos à limitação da liberdade, igualando as necessidades da coletividade. [Freitas and Castro 2013].

No Brasil, há um desafio do judiciário em determinar um tratamento constitucionalmente adequado ao discurso de ódio e garantir a manutenção do direito de livre expressão e pensamento. Outro problema é o enquadramento equivocado do discurso de ódio com algumas restrições à liberdade de expressão previstas em lei e passíveis de pena como crimes de ofensa à honra, calúnia, difamação e injúria.

Há uma tendência do Superior Tribunal Federal (STF) em não admitir o discurso de ódio nas redes sociais [Napolitano and Stroppa 2018] e basear decisões em normativas jurídicas com inspiração europeia, que equilibram liberdade de expressão com proteção, e restringido para algumas situações previstas em lei. No entanto, esta afirmação se afasta do entendimento aceito pela Suprema Corte Americana, que defende a liberdade de expressão de uma forma mais ampla em detrimento a outros direitos. Essa divergência de entendimento acentua a necessidade de uma discussão mais ampla da amplitude conferida a discurso de ódio.

Empresas de tecnologia sofrem pressão política por parte de usuários e de organizações para a aplicação de medidas que coíbam o discurso de ódio e não vinculem nomes e instituições a conteúdo ofensivo. Cada plataforma possui diretrizes próprias de restrição a discurso de ódio, definidas pela própria empresa. Por exemplo, *Facebook* [Facobook 2018] e *Twitter* [Twitter 2018?] monitoram denúncias por parte de usuário de postagens ligadas a conteúdo que incitam a violência ou promovem o ódio contra indivíduo ou grupo.

Devido à quantidade de dados *online* gerados diariamente, a classificação manual de discurso de ódio se torna uma tarefa de difícil execução. Além da compreensão de se a combinação de termos da sentença possa a vir a ser caracterizada como discurso de ódio, a criação manual de filtros a partir de palavras ou expressões pode não fornecer uma solução satisfatória de classificação. Através da classificação de textos e o uso de métodos de Aprendizado de Máquina é possível identificar discurso de ódio em documentos de forma automática e identificar se determinado documento pode ser considerado ou não um discurso de ódio.

2.2. Descoberta do Conhecimento

O aumento da capacidade de armazenamento de dados permitiu um rápido crescimento do volume de informação, excedendo a capacidade humana de compreensão de grandes conjuntos de dados. A análise de informação útil é uma tarefa custosa e subjetiva, sendo um desafio a execução manual de tal tarefa em grandes volumes de dados, tornando necessário o uso de técnicas computacionais e ferramentas que auxiliem no Processo da Descoberta de Conhecimento em Base de Dados.

O Processo de Descoberta de Conhecimento em Base de Dados (PDCBD), ou em inglês *Knowledge Discovery in Databases* (KDD), objetiva a aplicação de métodos e técnicas específicas para compreensão dos dados. Ele é definido como um processo não

trivial, interativo e iterativo para identificação de padrões válidos, potencialmente úteis e compreensíveis. [Fayyad et al. 1996b].

O PDCBD necessita de um usuário, que possui a capacidade de compreender as limitações e contribuições que o processo abrange, para que seja realizado o controle de uma sequência de etapas em busca de padrões que coincidam com os objetivos estabelecidos. Cada fase do PDCBD (Figura 1) possui características distintas, descritas a seguir:

(I) **Seleção de Dados**: o objetivo desta primeira etapa é o aprendizado do domínio de aplicação do problema, escolha de uma base apropriada para descoberta a ser realizada e decisão de qual será o modelo definido, baseado nos atributos e registros que irão compor a análise. Os dados de entrada podem possuir formatos distintos e podem ser selecionados de fontes de origem distintas [Tan et al. 2009];

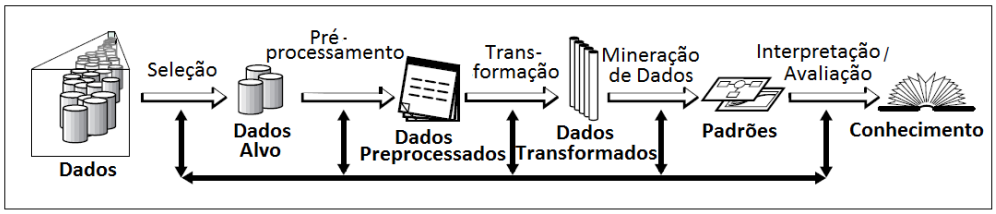
(II) **Pré-Processamento**: nesta etapa é realizada a adequação do dado bruto em um formato apropriado para análise e qualidade do resultado do processo. São aplicados métodos para identificação de dados redundantes, suavização de ruídos e remoção de informações ausentes ou incoerentes ao conjunto de dados selecionado [Fayyad et al. 1996a];

(III) **Transformação**: fase em que os dados obtidos na etapa anterior (etapa II) são processados, a fim de organizar informações consolidados de forma adequada aos objetivos do problema e aplicação do processo de DCBD em um formato conveniente para uma maior eficiência dos algoritmos a serem aplicados na etapa (IV) de Mineração de Dados [Goldschmidt and Passos 2005];

(IV) **Mineração de Dados**: É uma etapa multidisciplinar que combina métodos de análise de dados com algoritmos de processamento de dados, dentro do limite computacional. A escolha está fortemente associada ao conhecimento que se deseja extrair, com a exploração de padrões de interesse dos modelos de dados observados para análise e busca de conhecimento em potencial [Han et al. 2011];

(V) **Interpretação e Avaliação**: Nesta última etapa ocorre a interpretação e a descoberta de padrões a partir dos resultados obtidos, assim como a possibilidade de retorno para alguma das etapas anteriores. Também é realizada a visualização e interpretação dos padrões extraídos, traduzindo informações úteis em termos compreensíveis aos usuários para tomada de ações baseadas no conhecimento adquirido. [Fayyad et al. 1996b].

Figura 1. Etapas do Processo de Descoberta do Conhecimento



Fonte: Adaptado de [Fayyad et al. 1996b].

O Processo de Descoberta do Conhecimento é complexo e, apesar de utilizar técnicas computacionais que auxiliem na análise, é um processo cíclico, que necessita de participação humana para incorporação do conhecimento prévio adquirido e para garantia

de que o aprendizado é derivado dos dados. Caso os resultados não sejam satisfatórios ou não estejam consolidados conforme os objetivos do problema, poderão ser realizadas alterações durante o processo e iniciado um novo ciclo.

2.3. Análise de Sentimentos e Mineração de Opiniões

Com o crescimento de avaliações *online* e interação em redes sociais, há interesse de organizações em monitorar e compreender o comportamento de usuários em plataformas digitais. Extrair informações que expressam sentimentos positivos ou negativos de um documento é uma tarefa que exige compreensão da estrutura linguística e domínio do contexto do problema e a análise de Sentimento é uma área de pesquisa voltada para tal tarefa.

A Análise de Sentimentos, também chamada de Minerações de Opiniões, é definida como "o campo de estudo que analisa opiniões, sentimentos, avaliações, apreciações, atitudes e emoções de pessoas para com diferentes entidades como produtos, serviços, empresas, eventos problemas e seus atributos" [Liu 2012, p. 07]. Pesquisas da área de Análise de Sentimentos e Mineração de Opiniões objetivam o entendimento em variados documentos (ex. notícias, blogs, fóruns de discussão, etc) de qual é a intenção do comportamento de um usuário contida nas sentenças, expresso de forma racional ou emocional.

Um aspecto importante relacionado à Análise de Sentimentos é relativo à subjetividade ou objetividade de uma sentença e o impacto no processo de mineração de opiniões em um conjunto de dados. Uma sentença objetiva possui um fato ou informação que podem implicar em sentimentos positivos ou negativo [Feldman 2013]. Já uma sentença subjetiva expressa alguns sentimentos, opiniões e crenças pessoais que não necessariamente implicam em um sentimento positivo ou negativo, e, portanto, mais difíceis de detectar. [Liu 2012].

Pesquisas realizadas na área de Análise de Sentimentos e Mineração de Opiniões são baseadas na detecção do texto em diferentes granularidades em que a opinião pode ser processada em níveis distintos, sujeitos ao contexto e aplicação. Esta análise pode ser realizada em três níveis:

(I) **Nível de Documento**: a tarefa neste nível tem como objetivo classificar se o sentimento expresso em todo o documento é positivo ou negativo. Este nível de análise é aplicável a documentos que um texto é relacionado a um único assunto, como por exemplo, *review* de produtos [Pang et al. 2002];

(II) **Nível de Sentença**: a tarefa neste nível é relacionada a documentos que podem ser divididas em frases ou sentenças expressas individualmente como uma opinião positiva, negativa ou neutra (não significa opinião). Pesquisas também relacionam a classificação de subjetividade de cada sentença e que diferentemente de frases objetivas, podem não implicar em opiniões [Liu 2012];

(III) **Entidade e Nível de Aspecto**: O Nível de Aspecto é uma análise mais detalhada e o objetivo não é a estrutura de construção da linguagem e sim a própria opinião. Este nível de tarefa se baseia que uma opinião consiste em um sentimento (positivo ou negativo) e um alvo (opinião), transformando, desta forma, um texto não estruturado em estruturado. [Medhat et al. 2014].

Na etapa de Pré-Processamento é realizada a representação numérica para indicar

a presença ou ausência de determinados termos, através do emprego de técnicas mais simplificadas como Representação Binária e Representação de Frequência de Termos, ou mais complexa como a Representação TF-IDF.

Na Representação Binária um documento é representado por um vetor em que cada palavra (termo) é indicado pela sua presença ou ausência, por zeros e uns [Feldman and Sanger 2007]. Já na representação TF-IDF (*Term Frequency – Inverse Document Frequency*) é feita uma atribuição de peso de acordo com o número de ocorrências de um termo em um documento e a frequência deste termo em relação ao conjunto de documentos analisados. O objetivo é medir a relevância deste termo em uma coleção de documentos. [Liu 2012].

Pesquisas da Análise de Sentimentos também estabelecem como desafio a aplicação de métodos para extração de opiniões em textos em que os sentimentos não são expressos lexicalmente (ex. gírias ou ironia) e que possuem uma maior dificuldade de serem identificados, uma vez que são encontrados com menor frequência [Balahur et al. 2013] e pode alterar seu significado ao longo do tempo. Por ser um problema multifacetado, a Análise de Sentimentos utiliza métodos para auxiliar na tarefa de mineração de textos, que normalmente são divididos em duas classes: Aprendizado de Máquina, baseado em treinamento de um modelo previamente rotulado, e Métodos Léxicos, envolvendo identificação da polaridade a partir da orientação semântica das palavras contidas no texto. [Ravi and Ravi 2015].

2.4. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN), ou do inglês *Natural Language Processing* (NLP), objetiva o desenvolvimento de modelos computacionais que permitam executar tarefas que auxiliem no aprendizado e compreensão do conteúdo expresso em linguagem humana [Hirschberg and Manning 2015]. É uma etapa essencial de preparação dos dados, pois através de aplicações de técnicas para extração do texto analisado, é possível transformar o dado em um formato adequado e representativo para aplicação na etapa de Mineração de Textos.

Um desafio da área de PLN é o uso computacional de recursos linguísticos para processamento de palavras com propriedades distintas, inerentes ao idioma e contexto que assumem quando combinados entre si [Ranchhod 2001]. Para resolução do problema, a área de Processamento de Linguagem Natural envolve quatro etapas para entendimento da linguagem (Fig.2): análise morfológica (i), análise sintática (ii), análise semântica (iii) e análise pragmática (iv):

(I) **Análise Morfológica**: Trata em segmentar parágrafos ou frases em unidades mínimas rotuladas (ex. artigos, substantivos, verbos, etc.), denominados *tokens*, mantendo a semântica original do documento e constituindo uma espécie de dicionário. É possível ser realizado o processo de tokenização de textos através da utilização de delimitadores como pontuação, quebra de linhas, tabulações, alguns caracteres especiais e espaços contidos no texto [Liu 2012];

(II) **Análise Sintática**: A partir do *tokens* segmentados na Análise Morfológica é feita uma análise léxica buscando relacionamento entre as palavras, regras gramaticais e seu papel estrutural na construção de frases. Nesta etapa podem ser citados os métodos como remoção de *stopwords* (ex. de, para, com), que são palavras irrelevantes ao con-

texto da análise e *stemming*, técnica que identifica as variantes de uma palavra com uma representação comum com um mesmo significado para realização da remoção de seus prefixos e sufixos. Esta redução está ligada à língua de origem da palavra, pois depende das regras de formação da palavra [Hippisley 2010];

(III) **Análise Semântica**: Está relacionada ao significado de todo o conjunto resultante da sentença e do enunciado do contexto. Nesta etapa é pretendido o processamento da ambiguidade com a distinção de homônimos (diferentes significados para a mesma palavra), por exemplo "manga" (fruta) e 'manga" (blusa). Também são buscados diferentes sentidos para a mesma palavra, como "mão suja" e "passar a mão" [Goddard 2011];

(IV) **Análise Pragmática**: Busca a interpretação do texto como um todo e não mais de suas partes, buscando em outras sentenças a compreensão do texto que falta àquela frase em análise. É realizada a avaliação do que foi expresso em texto para determinar o seu real significado no contexto. [Dale et al. 2001].

Figura 2. Etapas do PNL

Fonte:Adaptado de [Dale et al. 2001].

2.5. Aprendizado de Máquina

Aprendizado de Máquina, também denominado *Machine Learning*, é uma área da Inteligência Artificial cujo o objetivo é o desenvolvimento de técnicas computacionais que auxiliem no processo de aquisição de conhecimento, identificação de padrões e tomada de decisão de forma automatizada. A vantagem de utilizar métodos de aprendizagem é a precisão comparável a alcançada por um especialista, uma vez que não há a necessidade de intervenção humana para a construção de um modelo ou portabilidade de domínio de aplicação. [Sebastiani 2002].

Algoritmos de classificação de dados visam identificar padrões de classes pré-definidas de um conjunto de dados de entrada, denominado conjunto de treinamento, de modo a encontrar um relacionamento entre atributos, predizendo a classe de um exemplo novo e desconhecido [Han et al. 2011]. Atualmente, existem inúmeras técnicas de Aprendizado de Máquina com características distintas e adequados para determinados problemas. Segundo [Russell and Norvig 2016, p. 740], métodos de classificação podem ser divididos em grupos, conforme sua necessidade de supervisão dos dados para treinamento: supervisionado, não supervisionado e aprendizagem por reforço.

No Aprendizado Supervisionado é fornecido ao algoritmo de aprendizado uma amostra de dados para treinamento, os quais apresentam um rótulo da classe associada para determinar padrões existentes nas classes da amostra. Já no aprendizado não supervisionado é executada a tarefa de treinamento sem o fornecimento de uma amostra de dados rotulados, sendo necessário que o próprio modelo identifique similaridades nos dados

para identificação de grupos de aprendizagem [Russell and Norvig 2016]. A Aprendizagem por reforço geralmente é utilizada em problemas em que as amostras de treinamento são escassas e custosas e caracterizado, por além de utilizar uma amostra de dados supervisionados, também utilizar um conjunto sem rótulo para treinamento [Ribeiro 2002]. Tendo em vista o problema de classificação textual abordado neste trabalho, serão abordados métodos de aprendizado supervisionado.

Redes Neurais Artificiais (RNA) são construções matemáticas inspiradas em modelos biológicos do sistema nervoso e que adquirem conhecimento através da experiência. A representação de uma rede neural envolve unidades de processamento (neurônios) altamente interconectadas de sinais de entrada recebidos por um conjunto de conexões, também denominado sinapses [Luger 2004]. A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados.

Considerando o problema de Minerações de Opiniões e Análise de Sentimentos, modelos clássicos da literatura são aplicados [Trivedi et al. 2015] como Regressão Logística (i) *Naive Bayes* (ii) e *Support Vector Machines* (iii), Árvore de Decisão (iv) e *Randon Forest* (v). Mais recentemente, pesquisas forma propostas utilizando arquiteturas de Redes Neurais [Dos Santos and Gatti 2014] como *Multilayer Perceptrons* (vi), *Convolutional Neural Network* (vii) e *Long Short-Term Memory Network* (viii). Estas técnicas citadas podem ser utilizadas em tarefas que envolvem Processamento de Linguagem Natural e auxiliar ao tema proposto no trabalho.

(I) **Regressão Logística (RL)**: um dos mais utilizados métodos de modelagem estatística de dados que objetiva descrever a relação de diversas variáveis independentes (categóricas ou não) e uma variável dependente binária, permitindo estimar diretamente a probabilidade de ocorrência de um determinado evento [Kleinbaum and Klein 2010];

(II) ***Naive Bayes* (NB)**: os classificadores que utilizam a técnica de estatística *Naive Bayes* são os mais utilizados em Mineração de Dados [Russell and Norvig 2016]. Essa técnica probabilística usa as frequências das ocorrências em uma base de dados para prever uma variável de interesse. É considerada simples ou ingênua (do inglês *naive*) por assumir que os atributos são condicionalmente independentes um dos outros para desta forma desconsiderar a correlação entre variáveis;

(III) ***Support Vector Machines* (SVM)**: é um método para classificação de dados lineares e não lineares, normalmente utilizado para problemas de reconhecimento. A classificação é realizada através do mapeamento de amostras de dados, a fim de identificar padrões pertencentes a diferentes classes do mesmo domínio de aprendizado [Han et al. 2011];

(III) ***Convolutional Neural Network* (CNN)**: são um tipo especializado de rede neural para processamento de dados projetados especificamente para reconhecimento de padrões. Em pelo menos uma de suas camadas, as mesmas empregam uma operação matemática chamada de convolução, um tipo especializado de operação linear onde as informações dos dados de entrada são processadas considerando campos receptivos locais. Adicionalmente inclui operações para reduzir a dimensionalidade espacial das representações produzindo uma saída [Goodfellow et al. 2016];

(IV) **Árvore de Decisão (AD)**: é um método adequado para categorização dos

dados organizado de forma hierárquica por nodos (nós) que são representação dos atributos e arestas direcionadas que recebem os possíveis valores de cada nó. Uma Árvore de Decisão utiliza o método de dividir para conquistar, ou seja, decompor um problema em subproblemas mais simples e aplicando recursivamente a mesma estratégia aplicada a cada subproblema [Tan et al. 2009];

(V) ***Random Forest*** **(RF)**: método baseado na agregação de múltiplas Árvores de Decisão contendo em cada árvore um subconjunto de atributos de entrada selecionados de forma aleatória a partir do modelo original. Cada árvore produz como saída uma classe predita, sendo selecionado como saída a classe com maior ocorrência ou média entre todas as árvores do modelo [Archer and Kimes 2008];

(VI) ***Multilayer Perceptrons*** **(MLP)**: são uma classe de redes neurais mais simples que possuem grande poder computacional e muito utilizadas para reconhecimento de padrões. Possuem uma ou mais camadas ocultas entre os nós computacionais de entrada e saída e não há conexões entre os neurônios de uma mesma camada, desta forma ocorrendo conexão apenas nas camadas seguintes [Bishop, 2006];

(VII) ***Convolutional Neural Network*** **(CNN)**: são um tipo especializado de rede neural para processamento de dados projetados especificamente para reconhecimento de padrões. Em pelo menos uma de suas camadas, as mesmas empregam uma operação matemática chamada de convolução, um tipo especializado de operação linear onde as informações dos dados de entrada são processadas considerando campos receptivos locais. Adicionalmente inclui operações para reduzir a dimensionalidade espacial das representações d produzindo uma saída [Goodfellow et al. 2016];

(VIII) ***Long Short-Term Memory Network*** **(LSTM)**: as redes LSTM conseguem preservar o erro e outros dados relevantes ao modelo através do tempo e das camadas. Ao manter um erro mais constante, eles permitem que redes recorrentes continuem aprendendo ao longo do tempo. [Tai et al. 2015].

2.6. Métodos de Medição de Performance e Validação

O uso de métricas de avaliação são utilizadas para determinar o desempenho de um classificador, calculado a partir dos resultados obtidos pelo algoritmo de classificação. Medidas comumente utilizadas para avaliação do modelo aplicado podem ser calculadas a partir de uma Matriz de Confusão que representa a frequência de classificação de cada classe do modelo. [Olson and Delen 2008].

Tabela 1. Matriz de Confusão de Duas Classes

Classe Correta	Classe Predita	
	Positivo	Negativo
Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte:Adaptado de [Olson and Delen 2008].

Conforme demonstrado na Tabela 1, um Verdadeiro Positivo (VP) demonstra o número de amostras positivas corretamente classificadas, Falso Positivo (FP) as amostras de outras classes erroneamente classificadas como positivo, Falso Negativo (FN)

a frequência de amostras rotuladas incorretamente como negativas e Verdadeiro Negativo (VN) a quantidade de amostras verdadeiramente classificadas como negativo [Han et al. 2011]. Medidas como Acurácia (i), *Recall* (ii), *Precision* (iii) e *F1-Score* (iv) e *Fleiss-Kappa* (v) são algumas métricas utilizadas para avaliação de um modelo e serão descritas a seguir:

(I) **Acurácia**: métrica que permite uma avaliação da proporção de acertos de um classificador, considerando tanto as amostras classificadas corretamente quanto incorretamente [Olson and Delen 2008];

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

(II) ***Recall***: também denominada de sensibilidade, é uma métrica de completude, que mede o total de acertos do classificador em relação à quantidade de amostras classificadas corretamente da classe alvo, adicionado a amostras que foram indevidamente atribuídas como negativas [Han et al. 2011];

$$Recall_{pos} = \frac{VP}{VP + FN} \tag{2}$$

$$Recall_{neg} = \frac{VN}{VN + FP} \tag{3}$$

(III) ***Precision***: é uma medida de exatidão que avalia a proporção de acertos do classificador considerando a classe alvo em relação tanto a amostras classificadas corretamente quanto pertencentes a outra classe [Han et al. 2011];

$$Precision_{pos} = \frac{VP}{VP + FP} \tag{4}$$

$$Precisionl_{neg} = \frac{VN}{VP + FN} \tag{5}$$

(IV) ***F1-Score***: também denominada *F-Measure* ou *F-Score*, é uma métrica que objetiva mostrar o balanço entre *Precision* e *Recall* [Goutte and Gaussier 2005];

$$\textit{F1-Score} = 2 * \frac{Precision * Recall}{precision + Recall} \tag{6}$$

(V) ***Fleiss-Kappa***: é uma variação da métrica *Cohen's Kappa* [Cohen 1960] e objetiva medir a confiabilidade do nível de concordância entre uma quantidade determinada de juízes quando avaliadas categorias não ordenadas de uma mesma amostra [Fleiss and Cohen 1973];

$$\textit{Fleiss-Kappa } (k) = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{7}$$

Além da utilização de métricas de avaliação para comparação dos resultados gerados entre dois ou mais classificadores, é recomendado o uso de métodos para determinar

a significância estatística dos resultados obtidos durante os experimentos. Para este objetivo, a validação cruzada (ou *k-fold cross validation*) é uma tradicional técnica utilizada para validar resultados obtidos e detectar a não generalização de um padrão. [Kim 2009].

O método de validação cruzada realiza o treinamento de um modelo particionando o conjunto de dados de entrada em k porções independentes, sendo uma destas partições retirado do conjunto de treinamento e utilizado como conjunto de testes. O desempenho é estimado como sendo o erro médio ou taxa de acerto média sobre estes k conjuntos de validação.

A medição de desempenho e avaliação dos resultados apresentados na etapa de treinamento é uma importante tarefa para avaliar a precisão de previsão do modelo e se o mesmo apresenta um bom desempenho em exemplos não utilizados durante o treinamento. No entanto, é importante a análise manual por parte de um especialista da métrica adequada à aplicação prática do modelo, pois os resultados são estatisticamente dependentes do problema e proporcional à quantidade de classes contidas no conjunto de dados.

3. Trabalhos Relacionados

Nesta seção são apresentadas pesquisas que aplicaram técnicas de Aprendizado de Máquina para identificação de discurso de ódio. Para encontrar trabalhos relevantes, foram pesquisados algumas variações de termos, conforme demonstrado na Tabela 2. As buscas foram realizadas através do *Google Scholar*[2], *Science Direct*[3], *ACM Digital Library*[4] e o Repositório Digital da Biblioteca Unisinos[5].

Tabela 2. Termos Pesquisados

Termos Pesquisados
machine learning
hate speech
machine learning hate speech
natural language processing
neural network hate speech
deep learning hate speech
RNA hate speech
liberdade expressão
discurso de ódio
aprendizado de máquina
aprendizado de máquina discurso de ódio
redes neurais
redes neuras discurso ódio
discurso ofensivo

Fonte: Elaborado pelo autor.

[2]Disponível em https://scholar.google.com.br
[3]Disponível em www.sciencedirect.com
[4]Disponível em https://dl.acm.org
[5]Disponível em http://www.unisinos.br/biblioteca

Com o crescimento do interesse na classificação automática e extração de opiniões, sentimentos e emoções contidas em texto, pesquisas em diversas subáreas da Inteligência Artificial e Processamento de Linguagem Natural vem nos últimos anos propondo métodos que auxiliem na identificação de sentimentos positivos e negativos [Pang et al. 2002, Riloff and Wiebe 2003] ou na classificação de sentenças subjetivas ou objetivas. [Yu and Hatzivassiloglou 2003, Riloff and Wiebe 2003].

3.1. Fonte dos Dados Coletados

Diversas fontes são utilizadas para extração de comentários *online* como sites, blogs e redes sociais. No trabalho de Xu e Zhu [2010] foi proposto um método para classificação de comentários da plataforma de vídeos *Youtube*[6] para substituição da primeira letra de cada palavra por asteriscos quando identificada como ofensivo. A classificação foi realizada em nível de sentença, utilizando a relação gramatical entre palavras, e baseado no método de Análise de Sentimentos proposto por Pang e Lee [2008].

Especificamente para o auxílio da identificação de discurso de ódio, os trabalhos apresentados até o momento realizaram experimentos extraindo comentários em plataformas *online* distintas. Warner e Hirschberg [2012] coletaram dados de grupos de notícias dos sites *Yahoo!* e *American Jewish Congress* para identificação de conteúdo antissemitas. Gilbert et al. [2018] extraíram as informações de um fórum de supremacia branca para a identificação de conteúdo odioso ou não odioso.

Outras plataformas como *Wikipedia*[7] [Wulczyn et al. 2017], *Reddit*[8] [Olteanu et al. 2018], *Facebook* [Del Vigna et al. 2017] e *Instagram*[9] [Liu et al. 2018] também já foram utilizadas em trabalhos anteriores. No entanto, trabalhos analisados majoritariamente realizam a extração de comentários da rede social *Twitter* [Pak and Paroubek 2010, Burnap and Williams 2015, Waseem and Hovy 2016], justificando a escolha em razão da facilidade de coleta de grandes volumes de dados que a plataforma proporciona. [Badjatiya et al. 2017].

Quanto ao idioma dos conjuntos de dados dos estudos analisados, a maioria foram originalmente publicados em língua inglesa. Entretanto, alguns trabalhos veem contribuindo nos últimos anos com pesquisas objetivando a classificação de discurso de ódio em outros idiomas. Ross et al. [2017] realizaram experimentos para identificação de discurso de ódio direcionado a refugiados, provenientes de comentários em alemão publicados na rede social Twitter. DelVigna et al. [2017] contribuíram com estudo de comentários no idioma italiano publicados em grupos e perfis extraídos do *Facebook*. Já Mubarak et.al [2017] propuseram um método de identificação de discurso de ódio utilizando comentários em árabe filtrados no *Twitter* a partir de palavras chaves analisadas como adequadas ao objetivo.

Outros trabalhos também propuseram soluções em outros idiomas como chinês [Su et al. 2017], holandês [Tulkens et al. 2016] e esloveno [Fišer et al. 2017]. O único trabalho disponível até o momento na língua portuguesa que propôs um método de identificação de discurso de ódio em português foi apresentado por Pelle e Moreira

[6]Disponível em www.youtube.com
[7]Disponível em www.wikipedia.org
[8]Disponível em www.reddit.com
[9]Disponível em www.instagram.com

[2017] com a extração de comentários de publicações de notícias das seções de esportes e políticas do site g1.globo.com[10].Posteriormente, os dados extraídos foram anotados manualmente, formando os conjuntos de dados que serão utilizados nos experimentos deste trabalho.

3.2. Anotação dos Dados

Após a extração dos dados, é importante compreender as limitações e de que forma poderão ser obtidos resultados satisfatórios ao objetivo da pesquisa durante a construção do modelo. Um processo utilizado é o emprego de anotadores humanos (do inglês *crowdsourcing*) para rotular manualmente os dados coletados de acordo com categorias estabelecidas para o objetivo do experimento e posteriormente serem utilizados como conjunto de dados para treinamento dos modelos de Aprendizado de Máquina. O uso deste método é justificado porque, além da compreensão de fatores como contexto social, gênero e localidade em que estão avaliadores, também mede o nível de concordância da classificação dos anotadores. [Warner and Hirschberg 2012].

Alguns trabalhos relataram o uso de anotadores humanos para categorização de comentários contendo ou não conteúdo ofensivo que posteriormente seriam aplicados nos experimentos propostos. Burnap e Williams [2015] utilizou um conjunto com 1901 comentários com rotulação binária (sim ou não) de discurso de ódio, anotados manualmente por participantes da pesquisa. Os dados anotados manualmente no trabalho de Waseem e Hovy [2016] formaram um conjunto com 16.000 comentários rotulados como racista, sexista ou neutro[11].

Pelle e Moreira [2017] desenvolveram uma aplicação para anotação manual de comentários classificados como não ofensivo ou ofensivo em categorias previamente definida (racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou xingamentos). A partir do resultado da anotação manual, os dados foram convertidos para uma categorização binária (sim ou não) para posterior aplicação dos classificadores.

Pesquisadores também usaram como recurso serviços para anotação manual dos dados, como *Hatebase.org*[12], que disponibiliza um dicionário de palavras ou expressões em vários idiomas para categorização ou não como discurso de ódio, anotados manualmente de forma voluntária por usuários da plataforma [Davidson et al. 2017, Zhang et al. 2018]. Outras ferramentas como *Figure Eight* [Burnap and Williams 2015][13] ou *Amazon Mechanical Turk* [Bigelow et al. 2016][14] permitem o carregamento de conjunto de dados para submissão de anotadores humanos que são recrutados pela plataforma.

3.3. Recursos de Processamento de Linguagem Natural

Um conjunto de dados em um formato adequado para aplicação dos modelos é uma etapa essencial para o experimento. A maioria dos trabalhos encontrados adaptam estratégias de Processamento de Linguagem Natural para auxiliar no problema de identificação

[10]Disponível em g1.globo.com
[11]Disponível em https://github.com/ZeerakW/hatespeech
[12]Disponível em https://hatebase.org/
[13]Anteriormente denominado CrowdFlower. Disponível em https://www.figure-eight.com/
[14]Disponível em https://www.mturk.com/

de discurso de ódio. Uma estratégia normalmente utilizada durante a etapa de pré-processamento é a normalização do conjunto de dados utilizando estratégias como conversão de caracteres em minúsculo [Burnap and Williams 2015], remoção de caracteres não alfabéticos [Pelle and Moreira 2017], urls [Nobata et al. 2016], *stopwords* [Liu and Forss 2015] e identificação de usuários. [Davidson et al. 2017].

O uso de dicionário é uma estratégia de mineração de textos para formação de uma lista de palavras para busca e contabilização de frequência de palavras [Bigelow et al. 2016] ou o uso desta abordagem com expressões regulares [Maloba 2014]. Outro recurso é a criação de um dicionário através da representação de palavras mais frequentes do conjunto de treinamento com o uso de técnicas como *Bag-of-Words* (BOW). [Greevy and Smeaton 2004, Kwok and Wang 2013, Burnap and Williams 2015].

A combinação de uma sequência de *n* palavras é uma forma de representação de palavras em atributos extraídos de textos [Badjatiya et al. 2017] e a técnica *N-Grams* é uma das mais utilizadas para detecção de discurso de ódio nos trabalhos relacionados, obtendo bons resultados quando combinados com outros recursos de PLN. [Liu and Forss 2015, Waseem and Hovy 2016, Davidson et al. 2017, Unsvåg and Gambäck 2018].

A técnica TF-IDF também foi utilizada para auxiliar na classificação de problemas através de vetores gerados a partir da medida de importância da palavra no corpus e sua frequência no documento [Dinakar et al. 2011, Liu and Forss 2015, Unsvåg and Gambäck 2018]. Outras técnicas relatadas em trabalhos anteriores são *Part-of-speech* (POS) para categorização gramatical de palavras em uma frase ou sentença [Del Vigna et al. 2017] e *Word Embeddings* (WE) para representação vetorial de palavras com significados semântico semelhantes em um documento. [Nobata et al. 2016, Badjatiya et al. 2017].

3.4. Classificadores

Por ser considerado um problema de classificação textual, o uso de técnicas supervisionadas de Aprendizado de Máquina é um recurso utilizado para identificação de discurso de ódio e amplamente encontrado na literatura. De acordo com pesquisa realizada por Fortuna e Nunes [2018] , *Support Vector Machines* (SVM) é o modelo mais utilizado para auxiliar na identificação de discurso de ódio. Outros classificadores utilizados são *Random Forest* [Burnap and Williams 2015], Árvore de Decisão [Davidson et al. 2017] [Davidson et al. 2017], Regressão Logística [Badjatiya et al. 2017] e *Naive Bayes*. [Pelle and Moreira 2017].

Nos últimos anos, alguns trabalhos também contribuíram com experimentos com modelos supervisionado de Redes Neurais Artificiais (RNA) explorando técnicas como *Multilayer Perceptron* (MLP) [Anzovino et al. 2018], *Convolutional Neural Network* (CNN) [Gambäck and Sikdar 2017], *Long Short-Term Memory Network* (LSTM) [Badjatiya et al. 2017] e *Gated Recurrent Unit Network* (GRU) [Zhang et al. 2018].

3.5. Avaliação dos Trabalhos Relacionados

As pesquisas apresentadas nesta seção estão relacionadas a este trabalho, seja pelo tema de identificação automática do discurso de ódio ou no emprego de técnicas que auxiliem na contribuição para a solução do problema. Na Tabela 3, os resultados dos trabalhos

relacionados são apresentados em ordem decrescente do ano de publicação. Foram selecionadas apenas as publicações que utilizaram ou informaram resultados dos métodos de Aprendizado de Máquina, pois alguns trabalhos tinham como foco de pesquisa apenas a análise semântica e morfológica.

Tabela 3. Avaliação dos Trabalhos Relacionados. A=Acurácia; P=Precision; R=Recall; F=F1-Score; ROC=Curva ROC

Ano	Trabalho	Fonte Dados	Idioma	Classes	Atributos	Técnica	Métricas	Melhor Resultado
2012	Warner and Hirschberg	Yahoo! e American Jewish Congress	Ing	sim ou não	POS	SVM	A, P, R, F	63% (F) - SVM
2013	Kwok and Wang	Twitter	Ing	sim ou não	BOW	NB	A	76% (A) - SVM
2015	Burnap and Williams	Twitter	Ing	sim ou não	BOW, N-gram	SVM, RL e RF	P, R, F	77% (F) - SVM
2015	Liu and Forss	N/I	Ing	multiclasse N/I	Stopwords	SVM, NB	P, R	53% (P) - SVM
2016	Tulkens et al	N/I	Ing	sim, não e neutro	BOW, N-gram, WE.	SVM	P, R, F, ROC	46% (F) - SVM
2016	Waseem and Hovy	Twitter	Ing	sim ou não	stop, NG	RL	P, R, F	74% (F) - RL
2017	Badjatiya et al	Twitter	Ing	racista, sexista, neutro	BOW, TF-IDF, N-gram, WE	CNN, LSTM, SVM	R, P, F	93% (F) - LSTM
2017	Davidson et al.	Twitter	Ing	sim ou não	POS, N-gram, TF-IDF	SVM, RL, NB, AD, RF	R, P, F	90% (F) - RL
2017	DelVigna et al	Facebook	Ita	sim ou não	POS, WE	SVM, LSTM	A, P, R, F	78% (F) LSTM
2017	Gambäck and Sikdar	Twitter	Ing	sim ou não	BOW, N-gram, WE	RL, CNN	P, R, F	78% (F) CNN
2017	Pelle and Moreira	G1	Pt-br	sim ou não	N-gram, FS	SVM, NB	F	81% (F) SVM
2017	Wulczyn et al.	Wikipedia	Ing	sim ou não	N-gram	RL, MLP, LSTM	A	96% (A) - MLP
2018	Gaydhani et. al	Twitter e Facebook	Ing	sexismo, racismo, nenhum	Stopwords, TFIDF	SMV, RL, NB	A, R, P, F	95% (F) RL
2018	Gilbert et al	Stormfront	Ing	sim ou não	BOW	SVM, CNN, LSTM	A	78% (A) LSTM
2018	Zhang, Robinson and Tepper	Yahoo!	Ing	ódio, ofensivo, neutro	N-gram	SVM, CNN, GRU	F	94% (F) - CNN e GRU

Fonte: Elaborado pelo autor.

A coluna “Ano” possui o ano de publicação do trabalho e a coluna “Trabalho” os autores do artigo. Já a coluna “Dataset” demonstra a origem dos dados utilizados durante os experimentos. “Idioma” detalha o idioma dos comentários extraídos para o conjunto. A coluna “Classe” informa os rótulos dos dados das publicações. A coluna “Atributo” informa os principais métodos de Processamento de Linguagem Natural utilizados nos trabalhos.

A coluna “Técnica” detalha os métodos de Aprendizado de Máquina utilizados nos trabalhos e “Métricas” informa os métodos de avaliação dos modelos utilizados. Por fim, a coluna “Melhor Resultado” informa, respectivamente, o melhor resultado apresentado, a métrica utilizada e o classificador. Não é possível concluir quais abordagens têm melhor desempenho, pois foram utilizados diferentes conjunto de dados e classes.

Entretanto, nota-se que vários autores utilizam métodos de Aprendizado de Máquina supervisionado.

Não foram encontrados trabalhos relacionados com a aplicação de modelos não supervisionados de Aprendizado de Máquina. Diversos métodos supervisionados foram aplicados com sucesso para resolução de problemas de classificação de textos e análise de sentimentos, e, que nos últimos anos vêm sendo utilizados com êxito para identificação de discurso de ódio em inglês. Por esta razão, este trabalho poderá contribuir com a análise de métodos de classificação supervisionados que obtiveram resultados satisfatórios para identificação de comentários ofensivos no idioma inglês e ainda não empregados em pesquisas no idioma português.

4. Materiais e Métodos

Este trabalho visa a identificação do discurso de ódio em comentários no idioma português em notícias selecionadas das seções de esportes e políticas e publicadas no site de notícias g1.com.br. O presente estudo avança o estado da arte em pesquisas de modelos computacionais e técnicas de Processamento de Linguagem Natural para classificação automática de discurso de ódio. Como diferencial deste trabalho, serão realizada análise de técnicas de classificação que obtiveram resultados satisfatórios para a identificação de conteúdo ofensivo no idioma inglês e que até o presente momento não foram utilizados em estudos no idioma português.

Nesta seção, são apresentados os métodos utilizados para pré-processamento e experimentos com técnicas de Aprendizado de Máquina, que possam vir a ser uma alternativa para identificação de discurso de ódio no idioma português. A subseção 4.1 descreve as principais características dos conjuntos de dados utilizados neste trabalho. Já a subseção 4.2 descreve as técnicas e algoritmos utilizados para implementação dos experimentos. A subseção 4.3 descreve as técnicas de classificação utilizadas neste trabalho. Este trabalho possui uma abordagem quantitativa, sendo utilizadas métricas para avaliação dos resultados dos experimentos realizados em conjunto de dados previamente classificados.

4.1. Conjunto de Dados

Para que seja alcançado o objetivo de investigar a utilização de técnicas de Aprendizado de Máquina para classificação de comentários online contendo discurso de ódio, foi utilizada a base proposta por Pele e Moreira [2017] e composta por comentários em português extraídos do site brasileiro de notícias g1.globo.com, delimitado às seções de esportes e política, selecionadas por haver maior ocorrências de comentários contendo conteúdo de ódio. Os dados possuem rotulação binária (ofensivos ou não ofensivos) e a classificação dos comentários foi realizada a partir de anotação manual.

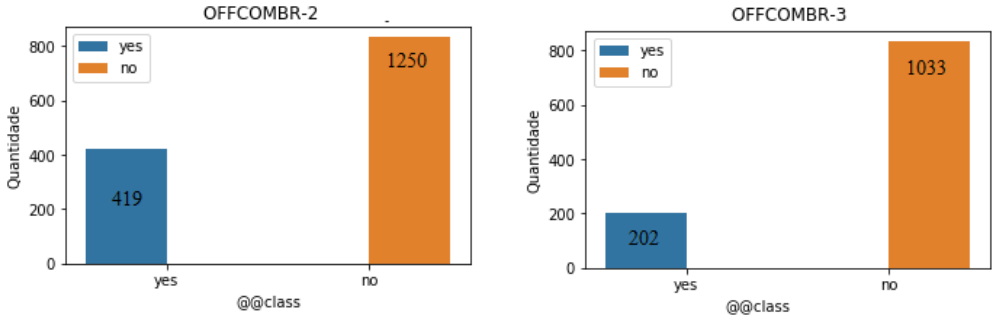
Os dados foram rotulados por sete participantes do experimento de Pelle e Moreira [2017], utilizando uma interface Web desenvolvida pelos autores[15], que exibia randomicamente os comentários a serem classificados. Os comentários que obtiveram concordância com dois ou mais anotadores resultou em dois conjuntos de dados, denominados *OFFCOMBR-2* e *OFFCOMBR-3*, e que serão descritos a seguir.

[15]Projeto disponível em https://github.com/rogersdepelle/hatedetector

O conjunto nomeado como *OFFCOMBR-2* é composto por 1250 comentários, sendo 419 rotulados como ofensivo (*yes*) ou não ofensivos (*no*) por no mínimo dois participantes. Para medir o nível de concordância entre os avaliadores do experimento, foi aplicado a métrica *Fleiss-Kappa* [Fleiss and Cohen 1973] que apresentou um coeficiente de concordância de 71%.

Já o conjunto *OFFCOMBR-3* possui 1033 comentários, destes 202 anotados manualmente como discurso de ódio e composto apenas por comentários que tiveram concordância da classificação como discurso de ódio ou não odioso por todos os três juízes participantes do experimento. Não foi utilizada a métrica Fleiss-Kappa, justificado por Pelle e Moreira (2017) não fazer sentido em razão deste conjunto ser composto apenas por anotações com 100% de concordância.

Figura 3. Estatísticas dos conjuntos *OFFCOMBR-2* e *OFFCOMBR-3*



Fonte:Elaborado pelo autor.

Os conjuntos de dados descritos são desbalanceados com amostras classificadas como não odiosas superando as amostras classificadas como discurso de ódio. Os comentários foram realizados no idioma português brasileiro com versões do corpus parcialmente pré processados por Pelle e Moreira [2017] , resultante da remoção de *emojis*, links de páginas web, identificação dos usuários que publicaram os comentários ou caracteres não alfabéticos.

Tabela 4. Atributos dos conjuntos de dados

Atributo	Descrição
Id	Código para identificação de cada linha.
@@class	Se contém ou não comentário ofensivo.
Document	Texto contendo comentários *online* de usuários coletados em sites de notícias.

Fonte: Elaborado pelo autor.

Para a leitura e manipulação dos dados a serem usados para treinamento e avaliação, os arquivos referentes aos conjuntos de dados *OFFCOMBR-2* e *OFFCOMBR-3*, originalmente disponibilizados no formato *Attribute-Relation File Format* (ARFF) [Witten et al. 2016], foram convertidos para o formato *Comma-Separated Values* (CSV) [Shafranovich 2005]. Já o atributo da classe contendo rótulo de discurso de ódio foi convertido para numérico, sendo atribuído o valor 1 para comentários ofensivos e 0 para não

ofensivo. A tabela 4 abaixo apresenta uma descrição dos três atributos dos conjuntos de dados *OFFCOMBR-2* e *OFFCOMBR-3*.

4.2. Especificações do Treinamento e da Avaliação

A plataforma utilizada para aplicação dos métodos apresentados nesta pesquisa foi o Python[16], linguagem open source que oferece uma série de pacotes que permitem o emprego de técnicas de Processamento de Linguagem Natural (PLN) e métodos de Aprendizado de Máquina, baseados no estado da arte de métodos de classificação e possibilitando a execução das tarefas do modelo proposto. O programa desenvolvido foi executado em uma máquina com um processador Intel *Core i7*, com *clock* de 1.8 GHz e 8 GB de memória. Para a leitura dos conjuntos de dados e execução dos métodos de classificação, foi utilizada a biblioteca para análise de dados *pandas*[17] em conjunto com as bibliotecas *numpy*[18] e *scipy*[19].

Inicialmente, foi realizada a preparação dos dados com remoção de *stopwords* da língua portuguesa utilizando o pacote *Natural Language Toolkit* [Loper and Bird 2002][20] com a adição manual de termos irrelevantes não presentes no dicionário utilizado e com maior ocorrência no conjunto de dados (ex. "kkk", "pra", "nessa"). Em seguida, foram criados novos conjuntos para os experimentos, com métodos distintos de PLN para cada um dos dois conjuntos de dados do trabalho a serem utilizados pelos classificadores do experimento propostos.

Foi adotado como padrão a atribuição de um prefixo, identificando as combinações de técnicas de PNL adotadas em cada conjunto utilizado durante os experimentos, sendo "Original" referentes aos comentários mantidos em sua forma original. Já os experimentos identificados com o prefixo "*Lower*" são comentários convertidos para minúsculo utilizando a biblioteca *Beautiful Soup*[21]. Também foram realizados experimentos com a aplicação da técnica *Bag-of-Words* (BOW) [Bespalov et al. 2011] com combinações de 01 a 03 termos nomeados com os sufixos "1G" (um termo), "2G" (um a dois termos) e "3G" (um a três termos).

Os conjuntos caracterizados como *Feature Selection* (FS) [Abbasi et al. 2008] contém apenas atributos com importância ao modelo maior que zero e está representado nos conjuntos com o sufixo "FS". Os métodos BOW e FS foram implementados utilizando a biblioteca *scikit-lear*[22], que disponibiliza uma série de ferramentas para Aprendizado de Máquina e também possibilitará a execução dos modelos de Aprendizado de Máquina propostos para este trabalho. A Tabela 5 demonstra o número de atributos selecionados em cada conjunto de dados utilizados para os experimentos deste trabalho.

A combinação de todos os métodos de Processamento de Linguagem Natural dos conjuntos *OFFCOMBR-2* e *OFFCOMBR-3* gerou 24 novos conjuntos com distinção de quantidade de atributos devido à redução de dimensionalidade após aplicação dos métodos de Processamento de Linguagem Natural mencionados anteriormente.

[16]Disponível em https://www.python.org
[17]Disponível em https://pandas.pydata.org
[18]Disponível em https://numpy.org
[19]Disponível em https://www.scipy.org
[20]Disponível em https://www.nltk.org
[21]Disponível em https://www.crummy.com/software/BeautifulSoup/bs4/doc
[22]Disponível em http://scikit-learn.org

Tabela 5. Número de Atributos dos Experimentos

Experimento	OFFCOMBR-2	OFFCOMBR-3
Lower_1G	3871	3397
Lower_2G	10502	9072
Lower_3G	16188	13963
Lower_1G_FS	1293	1112
Lower_2G_FS	1542	1326
Lower_3G_FS	1620	1397
Original_1G	4720	4091
Original_2G	12339	10580
Original_3G	18983	16248
Original_1G_FS	1405	1207
Original_2G_FS	1664	1419
Original_3G_FS	1750	1495

Fonte: Elaborado pelo autor.

4.3. Classificadores

Para treinamento dos conjuntos de dados, foram realizados experimentos com os classificadores Regressão Logística (RL) [Kleinbaum and Klein 2010], *Naive Bayes* (NB) [Russell and Norvig 2016], *Support Vetor Machine* (SVM) [Han et al. 2011], Árvore de Decisão (AD) [Tan et al. 2009] e *Random Forest* (RF) [Archer and Kimes 2008], utilizando a biblioteca *scikit-learn*. A escolha destes classificadores se justifica por serem amplamente utilizadas nos trabalhos relacionados para identificação de discurso de ódio no idioma inglês (com exceção de NB e SVM que também foram utilizados nos experimentos em português de Pelle e Moreira [2017]) e por não terem sido explorados para identificação de discurso de ódio em português.

Para o treinamento das Redes Neurais Artificiais (RNA), foi utilizada a biblioteca *keras*[23]. Para treinamento do conjunto, foram aplicados RNAs do tipo *Multilayer Perceptrons* (MLP) [Bishop 2006], *Convolutional Neural Network* (CNN) [Goodfellow et al. 2016] e *Long Short-Term Memory Network* (LSTM) [Tai et al. 2015]. A escolha destas técnicas foi fundamentada no estado da arte de métodos de classificação para a resolução do problema de identificação de discurso de ódio.

Foram realizados previamente alguns testes como, por exemplo, o tipo de hiperplano para SVM (linear ou rbf), critério de ponto de separação para AD (ganho de informação ou gini), número de estimadores para RF (50, 100, 150, 250, 500), tipo de regularização do modelo para RL (L1 ou L2), quantidade de épocas (3, 5, 12), camadas (1, 3, 5) e quantidade de amostras propagadas pela rede (32, 64, 128, 256) para as Redes Neurais Artificiais MLP, CNN e LSTM. Foram aplicadas alterações nas configurações dos classificadores nos casos em que apesentaram resultados relevantes, capacidade computacional disponíveis ou adequados aos modelos.

[23]Disponível em https://keras.io

Para *Support Vetor Machine* foi configurado o tipo de hiperplano linear para separar o dado e evitar *overfiting*. Para Árvore de Decisão foi mantido o padrão da biblioteca *sklearn* para medir a qualidade da divisão da árvore utilizando o índice Gini [Raileanu and Stoffel 2004]. Já para o *Random Forest*, foi definido em 250 o número de combinações de Árvores de Decisão do modelo. Para Regressão Logística, também foi mantida a configuração padrão da biblioteca *sklearn* de regularização L2 (*Ridge Regression*) [Owen 2007] para resolver ajustes excessivos do modelo e por ser eficiente em termos computacionais.

Para a execução das Redes Neurais Artificiais MLP, CNN e LSTM, foi utilizada uma rede *feedforward* [Hornik 1991] com somente uma camada escondida treinada através do algoritmo *backpropagation* [Hecht-Nielsen 1992], sendo definidas três épocas de aprendizado para adaptação dos pesos das redes neurais selecionadas e sendo 32 o número de amostras (*batch size*) propagadas para rede. Para compilação da rede, foi utilizado o método Adam [Kingma and Ba 2014] que realiza a otimização da taxa de aprendizado e função de perda da rede. O método Adam é considerado computacionalmente eficiente e que requer pouca memória, o que reduz o tempo de aprendizado.

Como métricas de avaliação dos resultados apresentados, foram utilizados Acurácia, *Precision*, *Recall* e *F1-Score*. Também foi utilizada a metodologia de validação cruzada *10-fold-cross-validation* [Kim 2009], sendo reservado para cada um dos 10 desdobramentos dos conjuntos de dados, 20% dos comentários para teste.

5. Apresentação e Análise dos Resultados

Esta seção apresenta os resultados da aplicação de técnicas de Aprendizado de Máquina Supervisionado sobre a versão da base de comentários de matérias do site de notícias g1.com.br de Pelle e Moreira [2017] . Foram realizados 12 experimentos em um contexto desbalanceado em cada um dos conjuntos de dados do experimentos utilizando 08 técnicas de classificação, totalizando 192 experimentos.

Os resultados apresentados são referentes à média obtida através das 10 execuções utilizando a metodologia de validação cruzada (*10-fold cross-validation*). Para a apresentação dos resultados, foram adotadas algumas abreviações: “POS” para identificar a classe classificada como positiva (não identificada como discurso de ódio) e “NEG” para a classe negativa, categorizada como discurso de ódio.

Para classificadores, foram adotadas as abreviações “AD” para identificar o classificador Árvore de Decisão, “CNN” para *Convolutional Neural Network,* “LSTM” para *Long Short Term Memory*, “MLP” para *Multilayer Perceptron*, “NB” para *Naive Bayes*, “RF” para *Randon Forest*, “RL” para Regressão Logística e “SVM” para *Support Vector Machine*.

Serão demonstrados os resultados para cada experimento, correspondente a diferentes combinações de técnicas de Processamento de Linguagem Natural. Em destaque em verde e laranja, os maiores e menores índices, respectivamente, de Acurácia, *Precion*, *Recall* e *F1-Score* para cada experimento e classificador aplicado. Em negrito, o melhor e pior resultado geral de cada métrica apresentada (Acurácia, *Precion*, *Recall* e *F1-Score*) com os experimentos dos conjuntos *OFFCOMBR-2* e *OFFCOMBR-3*.

5.1. Experimentos com *OFFCOMBR-2*

A Tabela 6 apresenta os resultados de Acurácia para cada experimento do conjunto *OFFCOMBR-2*. Pode ser percebido que os índices de Acurácia para CNN e LSTM apresentaram um empate estatístico em todos os experimentos, diferentemente do comportamento apresentados nos demais classificadores.

Tabela 6. Acurácia (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	64,20%	66,40%	66,40%	79,60%	70,80%	71,40%	73,50%	73,80%	70,76%
lower_1G_FS	61,70%	66,40%	66,40%	78,80%	70,80%	72,10%	73,50%	73,30%	70,38%
lower_1G+2G	63,60%	66,40%	66,40%	76,40%	71,40%	69,50%	73,20%	74,20%	70,14%
lower_1G+2G_FS	60,60%	66,40%	66,40%	79,20%	71,40%	71,50%	73,90%	73,30%	70,34%
lower_1G+2G+3G	62,60%	66,40%	66,40%	76,00%	71,40%	69,20%	73,10%	74,60%	69,96%
lower_1G+2G+3G_FS	60,60%	66,40%	66,40%	77,60%	71,40%	71,70%	73,90%	73,10%	70,14%
original_1G	61,10%	66,40%	66,40%	72,00%	69,40%	68,80%	71,60%	73,80%	68,69%
original_1G_FS	63,30%	66,40%	66,40%	74,80%	69,40%	70,10%	72,30%	71,10%	69,23%
original_1G+2G	61,40%	66,40%	66,40%	73,60%	70,10%	67,40%	71,50%	73,90%	68,84%
original_1G+2G_FS	63,50%	66,40%	66,40%	72,40%	70,10%	70,50%	72,60%	71,30%	69,15%
original_1G+2G+3G	62,80%	66,40%	66,40%	72,80%	70,10%	66,60%	71,70%	73,90%	68,84%
original_1G+2G+3G_FS	62,90%	66,40%	66,40%	71,20%	70,10%	70,90%	72,60%	71,30%	68,98%

Piores Resultados
Melhores Resultados

Fonte: Elaborado pelo autor.

O experimento *lower_1G* obteve melhores resultados para AD e MLP. Já *lower_1G_FS* apresentou melhor índice para RF, enquanto *lower_1G+2G* demonstrou melhor desempenho para NB. Os melhores resultados de *lower_1G+2G_FS* e *lower_1G+2G+3G_FS* foram em NB e RL e também ambos experimentos apresentaram o pior índice para AD. O experimento *lower_1G+2G+3G* obteve melhores resultados para NB e SVM.

Conforme os resultados apresentados, MLP foi o classificador que obteve o melhor índice geral de Acurácia (79,60%) do conjunto *OFFCOMBR-2* com o experimento lower_1G. O classificador MPL apresentou em geral os melhores resultados nos experimentos, exceto para *original_1G* e *original_1G+2G+3G* que apresentaram melhores índices para SVM, além de *original_1G+2G_FS* e *original_1G+2G+3G_FS* com melhor desempenho para RL. No entanto, vale destacar que mesmo que o resultado de MLP tenha ficado abaixo, o experimento *lower_1G* obteve o melhor resultado para Acurácia analisando a média obtida entre os classificadores com 71,10%, índice próximo a média inferior de original_1G (68,69%).

Analisando os resultados da Tabela 7 para *Recall*, pode ser verificado que *lower_1G* obteve os melhores resultados para MLP e RF e também um empate estatístico para SVM com *lower_1G+2G_FS*, *lower_1G+2G_FS* e *lower_1G+2G+3G_FS*. Já *lower_1G_FS* obteve os melhores resultados para RL e o melhor índice para NB juntamente com *lower_1G*.

Tabela 7. *Recall* (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	62,80%	66,40%	66,40%	**79,60%**	71,20%	72,00%	74,00%	76,40%	71,10%
lower_1G_FS	60,80%	66,40%	66,40%	78,80%	71,20%	70,80%	76,80%	76,00%	70,90%
lower_1G+2G	64,80%	66,40%	66,40%	76,40%	70,80%	69,20%	73,60%	74,80%	70,30%
lower_1G+2G_FS	61,20%	66,40%	66,40%	79,20%	70,80%	69,20%	76,40%	76,40%	70,75%
lower_1G+2G+3G	62,80%	66,40%	66,40%	76,00%	70,80%	68,00%	74,00%	74,80%	69,90%
lower_1G+2G+3G_FS	**60,40%**	66,40%	66,40%	77,60%	70,80%	68,80%	76,40%	76,40%	70,40%
original_1G	66,80%	66,40%	66,40%	72,00%	70,40%	68,40%	73,60%	70,40%	69,30%
original_1G_FS	62,00%	66,40%	66,40%	74,80%	70,40%	69,20%	73,20%	70,00%	69,05%
original_1G+2G	64,00%	66,40%	66,40%	73,60%	70,00%	68,00%	72,40%	72,40%	69,15%
original_1G+2G_FS	64,80%	66,40%	66,40%	72,40%	70,00%	68,00%	71,60%	70,40%	68,75%
original_1G+2G+3G	67,60%	66,40%	66,40%	72,80%	70,00%	67,20%	72,80%	72,40%	69,45%
original_1G+2G+3G_FS	63,60%	66,40%	66,40%	71,20%	70,00%	69,60%	71,60%	70,40%	68,65%

Fonte: Elaborado pelo autor.

O melhor resultado geral foi 79,60% de *Recall* para MLP com o experimento *lower_1G*, que também obteve a melhor média entre todos os classificadores utilizados nos experimentos. Os piores resultados foram obtidos com *original_1G+2G+3G_FS*, que apresentou os piores índices para MPL, NB e RL. O pior índice geral foi 60,40% de *Recall* com *lower_1G+2G+3G_FS* para AD.

A Tabela 8 demonstra os resultados dos algoritmos usando a métrica *Precision*, obtidos com os experimentos do conjunto *OFFCOMBR-2*. Avaliando os resultados, pode ser percebido melhores índices com o experimento *lower_1G_FS* para RL e MLP, sendo o último o melhor resultado geral do conjunto para *Precision* com o índice de 79,69%. O experimento *lower_1G* obteve melhor desempenho para RF. os experimentos com *lower_1G* e *lower_1G_FS* apresentaram o mesmo índice de 73,31 para NB. Já *original_1G+2G_FS* alcançou melhor resultado para AD e obteve o mesmo resultado com o experimento *lower_1G+2G+3G_FS* para SVM.

Tabela 8. *Precision* (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	71,16%	**44,09%**	**44,09%**	79,44%	73,31%	74,90%	74,37%	75,68%	67,13%
lower_1G_FS	71,49%	**44,09%**	**44,09%**	**79,69%**	73,31%	74,48%	76,38%	75,37%	**67,36%**
lower_1G+2G	73,06%	**44,09%**	**44,09%**	76,78%	72,80%	71,76%	74,68%	74,35%	66,45%
lower_1G+2G_FS	70,76%	**44,09%**	**44,09%**	78,71%	72,80%	72,61%	75,85%	75,90%	66,85%
lower_1G+2G+3G	70,34%	**44,09%**	**44,09%**	75,86%	72,80%	71,29%	76,18%	74,72%	66,17%
lower_1G+2G+3G_FS	71,79%	**44,09%**	**44,09%**	78,23%	72,80%	72,37%	75,85%	75,90%	66,89%
original_1G	72,18%	**44,09%**	**44,09%**	70,66%	71,58%	70,22%	73,62%	68,88%	64,42%
original_1G_FS	70,74%	**44,09%**	**44,09%**	75,20%	71,58%	71,76%	72,18%	69,58%	64,90%
original_1G+2G	69,89%	**44,09%**	**44,09%**	75,16%	71,31%	69,47%	73,77%	71,31%	64,89%
original_1G+2G_FS	75,98%	**44,09%**	**44,09%**	73,65%	71,31%	70,20%	70,13%	69,91%	64,92%
original_1G+2G+3G	74,50%	**44,09%**	**44,09%**	74,87%	71,31%	68,46%	75,60%	71,42%	65,54%
original_1G+2G+3G_FS	73,86%	**44,09%**	**44,09%**	69,93%	71,31%	71,51%	70,13%	69,91%	**64,35%**

Fonte: Elaborado pelo autor.

Analisando o desempenho geral dos experimentos para *Precision*, observa-se que os melhores resultados foram alcançados com *lower_1G_FS*. Ao avaliar o desempenho

dos classificadores, pode ser notado que MPL teve melhores índices entre os classificadores. Vale destacar que, em geral, os piores resultados ocorreram com os experimentos utilizando os comentários em seu formato original. Os classificadores com pior desempenho foram CNN e LSTM.

Tabela 9. *F1-Score* (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	63,66%	52,99%	52,99%	79,51%	71,81%	72,69%	70,56%	75,29%	67,44%
lower_1G_FS	61,40%	52,99%	52,99%	79,09%	71,81%	71,57%	75,24%	75,50%	67,57%
lower_1G+2G	65,63%	52,99%	52,99%	76,56%	71,40%	69,91%	69,47%	72,50%	66,43%
lower_1G+2G_FS	61,95%	52,99%	52,99%	78,68%	71,40%	69,99%	74,89%	76,04%	67,37%
lower_1G+2G+3G	63,73%	52,99%	52,99%	75,93%	71,40%	68,81%	69,52%	72,12%	65,94%
lower_1G+2G+3G_FS	60,90%	52,99%	52,99%	77,83%	71,40%	69,61%	74,89%	76,04%	67,08%
original_1G	67,72%	52,99%	52,99%	70,45%	70,83%	69,00%	70,23%	68,93%	65,39%
original_1G_FS	62,83%	52,99%	52,99%	74,97%	70,83%	69,91%	70,95%	69,76%	65,65%
original_1G+2G	64,99%	52,99%	52,99%	74,08%	70,47%	68,53%	67,34%	69,68%	65,13%
original_1G+2G_FS	65,34%	52,99%	52,99%	72,83%	70,47%	68,68%	69,41%	70,11%	65,35%
original_1G+2G+3G	68,45%	52,99%	52,99%	73,38%	70,47%	67,68%	67,33%	69,46%	65,34%
original_1G+2G+3G_FS	64,25%	52,99%	52,99%	70,07%	70,47%	70,20%	69,41%	70,11%	65,06%

Fonte: Elaborado pelo autor.

Os resultados para *F1-Score* apresentados na Tabela 9 demonstram que *lower_1G* obteve os melhores resultados para MPL, NB e RF. Já *lower_1G_FS* apresentou melhor desempenho para NB e RF, enquanto *lower_1G+2G_FS* e *lower_1G+2G+3G_FS* obtiveram os melhores índices para SVM. O experimento de *lower_1G+2G+3G_FS* teve melhor desempenho com SVM.

O experimento que apresentou o melhor desempenho para *F1-Score* foi *lower_1G_FS*, sendo o melhor índice de 79,09% para MLP. O classificador MLP também apresentou os melhores desempenhos entre os demais experimentos realizados no conjunto *OFFCOMBR-2*. O pior índice entre os experimentos foi apresentado por *original_1G+2G+3G* para NB, RF e RL. Os classificadores CNN e LSTM obtiveram os índices inferiores entre os experimentos obtendo o mesmo resultado de 52,99% em todos os experimentos.

Tabela 10. *F1-Score* POS (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM
lower_1G	66,18%	79,81%	79,81%	84,78%	76,92%	77,12%	82,85%	83,38%
lower_1G_FS	62,88%	79,81%	79,81%	83,49%	76,92%	75,75%	83,98%	82,56%
lower_1G+2G	68,12%	79,81%	79,81%	81,96%	76,68%	75,08%	82,81%	82,93%
lower_1G+2G_FS	63,94%	79,81%	79,81%	84,97%	76,68%	74,59%	83,66%	82,70%
lower_1G+2G+3G	66,67%	79,81%	79,81%	82,04%	76,68%	73,68%	83,20%	83,11%
lower_1G+2G+3G_FS	62,07%	79,81%	79,81%	82,72%	76,68%	74,17%	83,66%	82,70%
original_1G	71,48%	79,81%	79,81%	80,45%	76,88%	74,92%	82,54%	79,21%
original_1G_FS	65,20%	79,81%	79,81%	80,73%	76,88%	75,08%	81,74%	77,74%
original_1G+2G	68,75%	79,81%	79,81%	79,11%	76,49%	74,84%	82,26%	81,40%
original_1G+2G_FS	66,67%	79,81%	79,81%	78,37%	76,49%	74,36%	80,55%	78,11%
original_1G+2G+3G	71,38%	79,81%	79,81%	78,21%	76,49%	74,38%	82,65%	81,50%
original_1G+2G+3G_FS	65,92%	79,81%	79,81%	79,55%	76,49%	75,80%	80,55%	78,11%

Fonte: Elaborado pelo autor.

A Tabela 10 demonstra o resultado de F1-Score da classe positiva (sem discurso de ódio) com melhores índices obtidos com *lower_1G_FS*. Em geral, todos os classificadores

obtiveram bons resultados, sendo o melhor índice apresentado por RL. Vale destacar que os valores de MLP e SVN apresentaram resultados próximos aos de RL na predição da classe sem discurso de ódio.

Os resultados da Tabela 11 demonstram os índices para *F1-Score* da classe negativa (discurso de ódio). Os melhores resultados foram obtidos com os experimentos em que os comentários foram convertidos para minúsculo, sendo apresentados os melhores índices apresentados com *lower_1G_FS*.

Tabela 11. *F1-Score* NEG (Conjunto *OFFCOMBR-2*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM
lower_1G	58,67%	0,00%	0,00%	69,09%	61,70%	63,92%	46,28%	59,31%
lower_1G_FS	58,47%	0,00%	0,00%	**70,39%**	61,70%	63,32%	57,97%	61,54%
lower_1G+2G	60,71%	0,00%	0,00%	65,90%	60,96%	59,69%	43,10%	51,91%
lower_1G+2G_FS	58,01%	0,00%	0,00%	66,23%	60,96%	60,91%	57,55%	62,89%
lower_1G+2G+3G	57,92%	0,00%	0,00%	63,86%	60,96%	59,18%	42,48%	50,39%
lower_1G+2G+3G_FS	58,58%	0,00%	0,00%	68,18%	60,96%	60,61%	57,55%	62,89%
original_1G	60,29%	0,00%	0,00%	50,70%	58,89%	57,30%	45,90%	48,61%
original_1G_FS	58,15%	0,00%	0,00%	63,58%	58,89%	59,69%	49,62%	53,99%
original_1G+2G	57,55%	0,00%	0,00%	64,13%	58,56%	56,04%	37,84%	46,51%
original_1G+2G_FS	62,71%	0,00%	0,00%	61,88%	58,56%	57,45%	47,41%	54,32%
original_1G+2G+3G	62,67%	0,00%	0,00%	63,83%	58,56%	54,44%	37,04%	45,67%
original_1G+2G+3G_FS	60,94%	0,00%	0,00%	51,35%	58,56%	59,14%	47,41%	54,32%

Fonte: Elaborado pelo autor.

Em geral, todos os classificadores não obtiveram bons resultados para *F1-Score* NEG, sendo o pior demonstrado por RL. Também pode ser notado os piores resultados para CNN e LSTM para *F1-Score* da classe negativa, que em todos os experimentos com diferentes combinações de técnicas de Processamento de Linguagem Natural obtiveram índice zero. No entanto, vale destacar que melhores índices para MLP, SVM, NB e AD e RL apresentaram resultados próximos, mas se apresentaram na média dos índices alcançados.

5.2. Experimentos com *OFFCOMBR-3*

A Tabela 12 apresenta os resultados de Acurácia para cada experimento do conjunto *OFFCOMBR-3*. Pode ser percebido que os índices de Acurácia para CNN e LSTM apresentaram o mesmo padrão do conjunto *OFFCOMBR-2*, apresentando o mesmo resultado estatístico em todos os experimentos. Este fato deve ter ocorrido supostamente em razão do número de amostras da classe negativa e ocorrência de *overfiting* ao percorrer a rede.

O experimento *lower_1G* e *lower_1G_FS* obtiveram o mesmo resultado estatístico para NB. Já lower_1G+2G apresentou melhor desempenho para NB. Os melhores resultados de *original_1G* foram obtidos para MLP e SVM, enquanto o experimento *original_1G_FS* obteve o melhor índice para AD e RL. O melhor resultado ao analisar o índice médio entre todos os experimentos foi com *original_1G_FS*. Os piores índices foram de *original_1G+2G+3G*.

Tabela 12. Acurácia (Conjunto *OFFCOMBR-3*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	67,06%	76,33%	76,33%	81,64%	80,74%	84,62%	82,31%	83,16%	79,02%
lower_1G_FS	76,39%	76,33%	76,33%	81,16%	80,74%	80,99%	82,92%	82,08%	79,62%
lower_1G+2G	66,09%	76,33%	76,33%	80,19%	80,38%	84,86%	82,07%	83,04%	78,66%
lower_1G+2G_FS	75,55%	76,33%	76,33%	81,64%	80,38%	80,99%	82,92%	82,44%	79,57%
lower_1G+2G+3G	63,67%	76,33%	76,33%	81,16%	80,62%	84,01%	81,95%	83,16%	78,40%
lower_1G+2G+3G_FS	75,06%	76,33%	76,33%	79,71%	80,62%	81,11%	82,92%	82,44%	79,32%
original_1G	65,49%	76,33%	76,33%	82,61%	80,38%	83,65%	82,32%	**83,41%**	78,82%
original_1G_FS	76,88%	76,33%	76,33%	81,16%	80,38%	80,75%	83,05%	82,07%	79,62%
original_1G+2G	63,78%	76,33%	76,33%	81,64%	80,26%	83,28%	81,71%	83,05%	78,30%
original_1G+2G_FS	75,67%	76,33%	76,33%	82,13%	80,26%	79,91%	82,81%	82,07%	79,44%
original_1G+2G+3G	**63,43%**	76,33%	76,33%	80,68%	80,26%	83,40%	81,71%	82,80%	78,12%
original_1G+2G+3G_FS	76,51%	76,33%	76,33%	80,68%	80,26%	80,27%	82,81%	82,07%	79,41%

Piores Resultados
Melhores Resultados

Fonte: Elaborado pelo autor.

Conforme os resultados apresentados, SVM foi o classificador que obteve o melhor índice geral de Acurácia (83,41%), que, juntamente com RF, foram os classificadores com melhores índices dos experimentos com o conjunto *OFFCOMBR-3*. Pode ser também destacado que ao analisar isoladamente o resultado dos experimentos por classificador, poderá ser percebido uma estabilidade nos índices demonstrados.

Analisando os resultados da Tabela 13 para *Recall* Médio pode ser verificado que *lower_1G* obteve melhor resultado para NB. O experimento *lower_1G_FS* obteve os melhores índices para AD, RL, enquanto original_*1G* apresentou melhores índices para MLP e SVM. O experimento *original_1G+2G* obteve os melhores índices para RF. Pode ser destacado que NB apresentou o mesmo valor em todos os experimentos com comentários convertidos para minúsculo e o mesmo ocorrendo para RL com *lower_1G_FS*, *lower_1G+2G_FS* e *lower_1G+2G+3G_FS*.

Tabela 13. *Recall* (Conjunto *OFFCOMBR-3*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVN	Média
lower_1G	60,87%	76,33%	76,33%	81,64%	80,68%	80,19%	79,23%	80,19%	76,93%
lower_1G_FS	74,88%	76,33%	76,33%	81,16%	80,68%	77,29%	79,71%	78,26%	78,08%
lower_1G+2G	58,45%	76,33%	76,33%	80,19%	80,68%	80,68%	78,74%	80,68%	76,51%
lower_1G+2G_FS	73,43%	76,33%	76,33%	81,64%	80,68%	78,74%	79,71%	78,74%	78,20%
lower_1G+2G+3G	58,94%	76,33%	76,33%	81,16%	80,68%	79,71%	78,26%	80,19%	76,45%
lower_1G+2G+3G_FS	73,43%	76,33%	76,33%	79,71%	80,68%	78,74%	79,71%	78,74%	77,96%
original_1G	58,94%	76,33%	76,33%	82,61%	77,78%	80,68%	78,74%	82,13%	76,69%
original_1G_FS	71,01%	76,33%	76,33%	81,16%	77,78%	77,29%	79,23%	79,23%	77,30%
original_1G+2G	57,00%	76,33%	76,33%	81,64%	77,29%	81,16%	77,29%	80,68%	75,97%
original_1G+2G_FS	71,01%	76,33%	76,33%	82,13%	77,29%	76,33%	78,74%	80,68%	77,36%
original_1G+2G+3G	59,90%	76,33%	76,33%	80,68%	77,29%	79,71%	77,29%	81,16%	76,09%
original_1G+2G+3G_FS	71,01%	76,33%	76,33%	80,68%	77,29%	78,26%	78,74%	80,68%	77,42%

Fonte: Elaborado pelo autor.

O melhor resultado geral foi 82,61% de Recall para MLP com o experimento *original_1G*. No entanto, a melhor média entre todos os experimentos foi de

lower_1G+2G_FS. Em contraponto, os piores resultados foram obtidos em geral com *original_1G+2G*. O classificador que apresentou os melhores resultados foi MPL, mas outros classificadores apresentaram resultados semelhantes como NB, RL e SVM.

Tabela 14. *Precision* (Conjunto *OFFCOMBR-3*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM	Média
lower_1G	70,91%	**58,26%**	**58,26%**	80,37%	79,94%	78,61%	78,42%	78,44%	72,90%
lower_1G_FS	73,50%	**58,26%**	**58,26%**	80,20%	79,94%	75,18%	79,19%	75,54%	72,51%
lower_1G+2G	69,39%	**58,26%**	**58,26%**	78,33%	79,94%	80,56%	80,21%	80,56%	73,19%
lower_1G+2G_FS	72,13%	**58,26%**	**58,26%**	80,82%	79,94%	76,80%	79,19%	76,27%	72,71%
lower_1G+2G+3G	70,66%	**58,26%**	**58,26%**	80,20%	79,94%	79,19%	79,35%	79,90%	73,22%
lower_1G+2G+3G_FS	72,13%	**58,26%**	**58,26%**	77,91%	79,94%	76,80%	79,19%	76,27%	72,35%
original_1G	70,12%	**58,26%**	**58,26%**	**82,36%**	77,17%	79,56%	83,37%	82,32%	73,93%
original_1G_FS	69,78%	**58,26%**	**58,26%**	80,60%	77,17%	75,43%	79,38%	76,95%	71,98%
original_1G+2G	69,33%	**58,26%**	**58,26%**	81,21%	76,83%	84,89%	82,50%	82,66%	74,24%
original_1G+2G_FS	70,19%	**58,26%**	**58,26%**	82,32%	76,83%	74,05%	78,55%	78,96%	72,18%
original_1G+2G+3G	70,52%	**58,26%**	**58,26%**	79,54%	76,83%	81,56%	82,50%	84,89%	74,05%
original_1G+2G+3G_FS	70,19%	**58,26%**	**58,26%**	79,97%	76,83%	76,12%	78,55%	78,96%	72,14%

Fonte: Elaborado pelo autor.

A Tabela 14 demonstra o resultado de *Precision* obtidos durante os experimentos do conjunto *OFFCOMBR-3*. Avaliando os resultados, pode ser percebido os melhores resultados para o experimento *lower_1G_FS* para AD. O experimento *original_1G* obteve melhor desempenho para NB e MLP, sendo o último o melhor resultado geral do conjunto *OFFCOMBR-3*. Já *original_1G+2G* foi o experimento que alcançou melhor resultado para RF no conjunto *OFFCOMBR-3*. Vale destacar que em geral que os piores resultados ocorreram com os experimentos para os classificadores CNN e LSTM.

Tabela 15. *F1-Score* (Conjunto *OFFCOMBR-3*)

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVN	Média
lower_1G	63,72%	66,08%	66,08%	79,31%	**80,23%**	76,86%	74,02%	77,21%	72,94%
lower_1G_FS	74,07%	66,08%	66,08%	77,99%	**80,23%**	75,76%	74,86%	75,34%	73,80%
lower_1G+2G	61,53%	66,08%	66,08%	78,11%	**80,23%**	76,48%	72,02%	76,48%	72,13%
lower_1G+2G_FS	72,70%	66,08%	66,08%	78,72%	**80,23%**	77,18%	74,86%	76,05%	73,99%
lower_1G+2G+3G	62,00%	66,08%	66,08%	77,99%	**80,23%**	74,86%	71,06%	75,68%	71,75%
lower_1G+2G+3G_FS	72,70%	66,08%	66,08%	76,11%	**80,23%**	77,18%	74,86%	76,05%	73,66%
original_1G	61,99%	66,08%	66,08%	79,84%	77,44%	77,25%	71,38%	78,79%	72,36%
original_1G_FS	70,34%	66,08%	66,08%	77,64%	77,44%	76,01%	73,50%	76,09%	72,90%
original_1G+2G	**60,23%**	66,08%	66,08%	78,39%	77,04%	75,97%	68,30%	75,59%	70,96%
original_1G+2G_FS	70,58%	66,08%	66,08%	78,79%	77,04%	74,73%	72,61%	78,51%	73,05%
original_1G+2G+3G	62,86%	66,08%	66,08%	79,87%	77,04%	73,85%	68,30%	75,97%	71,26%
original_1G+2G+3G_FS	70,58%	66,08%	66,08%	76,88%	77,04%	76,53%	72,61%	78,51%	73,04%

Fonte: Elaborado pelo autor.

Os resultados para *F1-Score* apresentados na Tabela 15 demonstram que *lower_1G_FS* obteve os melhores resultados para AD e RL. Já *lower_1G+2G_FS* apresentou melhor desempenho geral entre os experimentos e para os classificadores NB, RL. O experimento *original_1G* foi o melhor desempenho para RF e *original_1G+2G+3G* para MLP. O classificador NB obteve o melhor resultado geral, sendo o resultado de 60,23% apresentados em todos os experimentos com *lower* para este classificador.

Em contraponto, os experimentos realizados com comentários em seu formato original, em geral, apresentaram os piores índices do conjunto*OFFCOMBR-3*, sendo original_1G+2G_com o pior resultado geral e também o índice médio inferior do conjunto.

Tabela 16. *F1-Score* POS Conjunto *OFFCOMBR-3*

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM
lower_1G	70,55%	86,58%	86,58%	88,82%	87,58%	88,12%	87,82%	88,05%
lower_1G_FS	83,95%	86,58%	86,58%	88,70%	87,58%	85,80%	88,07%	86,80%
lower_1G+2G	68,38%	86,58%	86,58%	87,83%	87,58%	88,57%	87,71%	88,57%
lower_1G+2G_FS	82,97%	86,58%	86,58%	88,95%	87,58%	86,75%	88,07%	87,06%
lower_1G+2G+3G	68,40%	86,58%	86,58%	88,70%	87,58%	88,07%	87,47%	88,32%
lower_1G+2G+3G_FS	82,97%	86,58%	86,58%	87,86%	87,58%	86,75%	88,07%	87,06%
original_1G	68,63%	86,58%	86,58%	**89,53%**	85,63%	88,44%	87,78%	89,34%
original_1G_FS	81,37%	86,58%	86,58%	88,76%	85,63%	85,71%	87,89%	87,46%
original_1G+2G	**66,67%**	86,58%	86,58%	89,02%	85,27%	89,01%	87,05%	88,70%
original_1G+2G_FS	81,25%	86,58%	86,58%	89,34%	85,27%	85,20%	87,64%	88,17%
original_1G+2G+3G	69,60%	86,58%	86,58%	87,73%	85,27%	88,20%	87,05%	89,01%
original_1G+2G+3G_FS	81,25%	86,58%	86,58%	88,51%	85,27%	86,49%	87,64%	88,17%

Fonte: Elaborado pelo autor.

A Tabela 16 demonstra o resultado de *F1-Score* da classe positiva (sem discurso de ódio) com melhores índices obtidos com lower_1G+2G_S. Em geral, todos os classificadores obtiveram bons resultados, sendo o melhor resultado apresentado por MLP. Vale destacar que os valores de RF, NB e SVN apresentaram resultados próximos aos índices de MLP. Em geral, o melhor desempenho ocorreu com os experimentos com comentários convertidos em minúsculo, exceto original_1G+2G, que obteve o melhor desempenho para RF. O mesmo experimento obteve o pior desempenho médio e também o pior índice geral para AD.

Tabela 17. *F1-Score* NEG Conjunto *OFFCOMBR-3*

Experimento	AD	CNN	LSTM	MLP	NB	RF	RL	SVM
lower_1G	41,73%	**0,00%**	**0,00%**	48,65%	56,52%	40,58%	29,51%	42,25%
lower_1G_FS	42,22%	**0,00%**	**0,00%**	43,48%	56,52%	43,37%	32,26%	38,36%
lower_1G+2G	39,44%	**0,00%**	**0,00%**	46,75%	56,52%	37,50%	21,43%	37,50%
lower_1G+2G_FS	39,56%	**0,00%**	**0,00%**	45,71%	56,52%	46,34%	32,26%	40,54%
lower_1G+2G+3G	41,38%	**0,00%**	**0,00%**	43,48%	56,52%	32,26%	18,18%	34,92%
lower_1G+2G+3G_FS	39,56%	**0,00%**	**0,00%**	38,24%	56,52%	46,34%	32,26%	40,54%
original_1G	40,56%	**0,00%**	**0,00%**	48,57%	51,06%	41,18%	18,52%	44,78%
original_1G_FS	34,78%	**0,00%**	**0,00%**	41,79%	51,06%	44,71%	27,12%	39,44%
original_1G+2G	39,46%	**0,00%**	**0,00%**	44,12%	50,53%	33,90%	7,84%	33,33%
original_1G+2G_FS	36,17%	**0,00%**	**0,00%**	44,78%	50,53%	40,96%	24,14%	47,37%
original_1G+2G+3G	41,13%	**0,00%**	**0,00%**	**54,55%**	50,53%	27,59%	7,84%	33,90%
original_1G+2G+3G_FS	36,17%	**0,00%**	**0,00%**	39,39%	50,53%	44,44%	24,14%	47,37%

Fonte: Elaborado pelo autor.

Os resultados da Tabela 17 demonstram que para *F1-Score* da classe negativa (discurso de ódio) todos os experimentos obtiveram índices abaixo de 60%. O melhor

resultado geral para *F1-Score* NEG do conjunto *OFFCOMBR-3* foi apresentado com *original_1G+2G+3G* e o melhor resultado médio entre os classificadores foi com o experimento *original_1G+2G_FS*.

Os classificadores CNN e LSTM apresentaram índice zero para *F1-Score* NEG com todas as combinações de técnicas de Processamento de Linguagem Natural utilizados nos experimentos do conjunto *OFFCOMBR-3*, assim como ocorrido com o conjunto *OFFCOMBR-2*. Supostamente, isso se deve ao número de amostras da classe negativa e ocorrência de *overfiting* ao percorrer a rede.

6. Discussão

Considerando isoladamente as duas variações do conjunto *OFFCOMBR*, foram obtidos alguns resultados satisfatórios com a aplicação de modelos de classificação para identificação de discurso de ódio em português em um contexto desbalanceado. Ao avaliar os resultados dos conjuntos *OFFCOMBR-2* e *OFFCOMBR-3*, pode ser percebido que os índices para *F1-Score* demonstraram um balanceamento com os resultados para *Recall* e *Precision*.

A Tabela 18 demonstra o melhor índice geral para cada classificador e em qual experimento obteve o resultado. O melhor desempenho do conjunto *OFFCOMBR-2* foi MPL para *F1-Score* de 79,51% e *OFFCOMBR-3* obteve 80,23% para NB. O pior índice de *OFFCOMBR-2 F1-Score* foi 52,99% para CNN e LSTM e 66,08% para CNN e LSTM em *OFFCOMBR-3*.

Tabela 18. Melhores Resultados para *F1-Score*

	OFFCOMBR-2		OFFCOMBR3	
Técnica	F1-Score	Experimento	F1-Score	Experimento
AD	68,45%	*original_1G+2G+3G*	74,07%	*lower_1G_FS*
CNN	52,99%	todos	66,08%	todos
LSTM	52,99%	todos	66,08%	todos
MLP	**79,51%**	*lower_1G*	79,87%	*original_1G+2G+3G*
NB	71,81%	*lower_1G e lower_1G_FS*	**80,23%**	todos *lower*
RF	72,69%	*lower_1G*	77,25%	*original_1G*
RL	75,24%	*lower_1G_FS*	74,86%	*lower_1G_FS, lower_1G+2G_FS e lower_1G+2G+3G_FS*
SVN	76,04%	*lower_1G+2G_FS e lower_1G+2G+3G_FS*	78,79%	*original_1G*

Fonte: Elaborado pelo autor.

Conforme demosntrado na Tabela 19, o classificador com melhor desempenho de *OFFCOMBR-2* foi MLP com índice médio para *F1-score* de 75,28%, com resultado 6% superior ao melhor desempenho médio apresentado por Pelle e Moreira [2017] para SVM. Os resultados de NB e SVM também experimentados neste trabalho confirmam índices apresentados no trabalho de referência. Apesar de apresentarem valores inferiores em relação a MPL, vale destacar os índices de RL e RF ficaram muito próximos ao melhor desempenho médio entre os classificadores.

Tabela 19. Média experimentos para *F1-Score OFFCOMBR-2* e *OFFCOMBR-3*

Técnica	OFFCOMBR-2	OFFCOMBR3
AD	64,24%	66,94%
CNN	52,99%	66,08%
LSTM	52,99%	66,08%
MLP	**75,28%**	78,30%
NB	71,06%	**78,70%**
RF	69,72%	76,06%
RL	70,77%	72,37%
SVN	72,13%	76,69%

Fonte: Elaborado pelo autor.

A partir dos resultados apresentados por este trabalho, pode ser percebido que não houve melhoras significantes ao reduzir a dimensionalidade dos atributos do conjunto de dados com o uso das técnicas *Bag-of-Words* e *Feature Selection* (FS). Pode ser também constatado que ao comparar a mesma combinação de técnicas de PLN com comentários em seu formato original, a conversão para minúsculo promoveu uma melhora dos resultados para *OFFCOMBR-2*.

Como esperado, em razão de *OFFCOMBR-3* ser formado a partir de três anotadores manuais em concordância, o desempenho deste conjunto foi melhor em relação a *OFFCOM-2*. O melhor resultado médio entre os classificadores foi obtido com a NB com 78,70%, índice muito próximo a MLP (78,30%). Ao analisar os experimentos realizados, pode ser percebido melhoras nos índices tanto ao aplicar o método *Feature Selection* (FS) para redução de dimensionalidade, assim como com a conversão para minúsculo dos comentários utilizados durante a etapa aplicação de modelos de classificação. No entanto, diferentemente do esperado, os experimentos com *Bag-of-Words* não apresentaram melhoras nos índices quando reduzida a dimensionalidade do conjunto com o agrupamento de dois ou mais atributos.

Também pode ser destacado os baixos índices para *F1-Score* para a classe negativa (discurso de ódio), com o melhor índice apresentado de 56,52%. Os resultados deste trabalho apresentam o uso de RNA *Multilayer Perceptron* e *Random Forest* como alternativas para identificação de padrões de linguagem contendo discurso de ódio em português e também reforçaram os resultados apresentados Pelle e Moreira [2017]. A comparação com o trabalho de referência foi parcial em razão do mesmo apresentar apenas os índices médios de *F1-score*. A comparação com outros trabalhos não é possível, pois os mesmos utilizaram outros conjuntos de dados e idiomas em seus experimentos.

7. Conclusão

Este trabalho realizou um estudo investigativo de técnicas de Aprendizado de Máquina quanto à viabilidade de utilização de possíveis métodos de classificação como alternativa para melhorar a tarefa de identificação de discursos de ódio em comentários textuais online. Embora o assunto seja amplamente discutido na literatura, sobretudo em inglês, até o presente momento da realização deste trabalho, foram encontrados na literatura apenas

o estudo de Pelle e Moreira [2017] para resolução do problema utilizando comentários em português e utilizado como referência para este trabalho. Apesar de outras pesquisas terem apresentado resultados satisfatórios com outros métodos de classificação para a língua inglesa, apenas *Support Vector Machines* e *Naive Bayes* foram utilizados para português [Pelle and Moreira 2017].

Para tal objetivo, foi utilizada uma versão das bases *OFFCOMBR-2* e *OFFCOMBR-3* de Pelle e Moreira [2017] , contendo comentários em português e publicados no site de notícias g1.com.br. Foram utilizadas diferentes combinações de técnicas de Processamento de Linguagem Natural (conversão comentários para minúsculo ou formato original, *Bag-of-Words*, *Feature Selection*) para analisar diferentes cenários ao aplicar diferentes classificadores (NB, SVM, AD, RF,RL, MLP, CNN, LSTM). É destacado que o treinamento foi realizado em conjuntos de dados desbalanceados, contendo amostras classificadas como não odiosas superando a classificadas como discurso de ódio, o que ficou evidente ao serem analisados os resultados para *F1-Score* NEG.

O uso da métrica *F1-Score* foi determinante para realizar um comparativo com os resultados de Pelle e Moreira [2017], que realizaram seus experimentos com os classificadores NB e SVM. Os resultados do conjunto *OFFCOMBR-2* com o classificador MLP (79%) foram superiores aos apresentados nos experimentos de Pelle e Moreira [2017] para SVM (77%) e NB(71%). Já para *OFFCOMBR-3*, os mesmos classificadores utilizados por Pelle e Moreira [2017] obtiveram melhor desempenho; no entanto, MLP (80%) e RF (77%) obtiveram desempenho semelhantes a SVM (82%) e NB (79%) do trabalho de referência.

Em geral, os resultados obtidos com as combinações de métodos PNL e o treinamento das amostras com os classificadores MLP, SVM, RF e NB foram satisfatórios, devido ao fato destes classificadores serem técnicas amplamente utilizadas para classificação de textos e especialmente no campo de Análise de Sentimentos.

Para trabalhos futuros é sugerido:

- Aplicação de métodos para balanceamento para geração de amostras sintéticas do conjunto de dados para posterior aplicação de métodos de classificadores;
- Formação de um conjunto de treinamento com mais amostras para validação dos modelos aplicados;
- Aplicação de classificadores em outros conjuntos de dados no idioma português;
- Explorar outras configurações dos experimentos apresentado com a combinação de *word embeddings* em português para aprimoramento dos resultados.

Referências

Facebook community standards. Disponível em: https://www.facebook.com/communitystandards/hate_speech. Acesso em: 29 maio 2019.

Indicadores de denúncias realizadas entre 2006 e 2018. Disponível em: http://indicadores.safernet.org.br/. Acesso em: 08 maio 2019.

Twitter hateful conduct policy. Disponível em: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Acesso em: 29 maio 2019.

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2013). Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375–382. ACM.

Bigelow, J. L., Edwards, L., et al. (2016). Detecting cyberbullying using latent semantic indexing. In *Proceedings of the First International Workshop on Computational Methods for CyberSafety*, pages 11–14. ACM.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Brugger, W. (2007). Proibição ou proteção do discurso do ódio? algumas observações sobre o direito alemão e o americano. *Revista de Direito Público*.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

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

Dale, R., Moisl, H., and Somers, H. (2001). Handbook of natural language processing.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Del Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.

Dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

Dresch, A., Lacerda, D. P., and Júnior, J. A. V. A. (2015). *Design science research: método de pesquisa para avanço da ciência e tecnologia*. Bookman Editora.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996b). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Feldman, R. and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Freitas, R. S. d. and Castro, M. F. d. (2013). Liberdade de expressão e discurso do ódio: um exame sobre as possíveis limitações à liberdade de expressão. *Sequência (Florianópolis)*, pages 327–355.

G1. Investigação policial conclui que morte de moa do katendê foi motivada por briga política. Disponível em: https://g1.globo.com/ba/bahia/noticia/2018/10/17/investigacao-policial-conclui-que-morte-de-moa-do-katende-foi-motivada-por-briga-politica-inquerito-foi-enviado-ao-mp.ghtml. Acesso em: 12 out. 2019.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Georgescu, M. (2014). *Bookmarks: A manual for combating hate speech online through human rights education*. Council of Europe.

Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Goddard, C. (2011). *Semantic analysis: A practical introduction*. Oxford University Press.

Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático*. Gulf Professional Publishing.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer.

Greevy, E. and Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier.

Hippisley, A. R. (2010). Lexical analysis.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Horbach, B. B. (2012). Os limites da liberdade de expressão. *Revista Brasileira de Direitos Fundamentais & Justiça*, 6(20):218–235.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinbaum, D. G. and Klein, M. (2010). Introduction to logistic regression. In *Logistic regression*, pages 1–39. Springer.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Liu, P., Guberman, J., Hemphill, L., and Culotta, A. (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth International AAAI Conference on Web and Social Media*.

Liu, S. and Forss, T. (2015). New classification models for detecting hate and violence web content. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 487–495. IEEE.

Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Luger, G. F. (2004). *Inteligência Artificial-: Estruturas e estratégias para a solução de problemas complexos*. Bookman.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Maloba, W. J. (2014). *Use of regular expressions for multi-lingual detection of hate speech in Kenya*. PhD thesis, iLabAfrica.

Martins, E. Bar de refugiados em sp lamenta ataque: 'crescente discurso de intolerância e ódio'. Disponível em: https://oglobo.globo.com/mundo/bar-de-refugiados-em-sp-lamenta-ataque-crescente-discurso-de-intolerancia-odio-23920405. Acesso em: 19 de out. de 2019.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Napolitano, C. J. and Stroppa, T. (2018). O supremo tribunal federal e o discurso de ódio nas redes sociais: exercício de direito versus limites à liberdade de expressão. *Revista Brasileira de Políticas Públicas*, 7(3):313–332.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Nogueira, E. Estupro em fortaleza pode ter ocorrido por intolerância política. Disponível em: http://agenciabrasil.ebc.com.br/politica/noticia/2018-10/estupro-em-fortaleza-pode-ter-ocorrido-por-intolerancia-politica. Acesso em: 12 de out. de 2019.

Olson, D. L. and Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Olteanu, A., Castillo, C., Boy, J., and Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443(7):59–72.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*. SBC.

Raileanu, L. E. and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93.

Ranchhod, E. M. (2001). *Tratamento das línguas por computador: uma introdução à linguística computacional e suas aplicações*.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Ribeiro, C. (2002). Reinforcement learning agents. *Artificial intelligence review*, 17(3):223–250.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Rothenburg, W. C. and Stroppa, T. (2015). Liberdade de expressão e discurso de ódio: o conflito discursivo nas redes sociais. In *3º Congresso Internacional de direito e contemporaneidade*.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Sarmento, D. (2006). A liberdade de expressão e o problema do "hate speech". *SARMENTO, Daniel. Livres e iguais: estudos de direito constitucional. Rio de Janeiro: lúmen juris*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Shafranovich, Y. (2005). Common format and mime type for comma-separated values (csv) files.

Su, H.-P., Huang, Z.-J., Chang, H.-T., and Lin, C.-J. (2017). Rephrasing profanity in chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pages 18–24.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao datamining: mineração de dados*. Ciência Moderna.

Trivedi, M., Sharma, S., Soni, N., and Nair, S. (2015). Comparison of text classification algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 4(02).

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Unsvåg, E. F. and Gambäck, B. (2018). The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Xu, Z. and Zhu, S. (2010). Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pages 1–10.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.