Universidade Federal de Alagoas

Instituto de Computação

Coordenação de Pós-Graduação em Informática

Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos

Sebastião Rogério da Silva Neto

Proposta de Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal de Alagoas.

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Computação Visual e Inteligente

Evandro de Barros Costa

Rafael Ferreira Leite de Mello

(Orientadores)

Maceió, Alagoas, Brasil

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Bibliotecária Responsável: Janaina Xisto de Barros Lima

S586u Silva Neto, Sebastião Rogério da.
Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos / Sebastião Rogério da Silva Neto. – 2017.
95 f. : il.

Orientador: Evandro de Barros Costa.
Coorientador: Rafael Ferreira de Leite Mello.
Dissertação (mestrado em Informática) - Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió, 2017.

Bibliografia: f. 75-80.
Apêndice: f. 82-95.

1. Mineração de texto (Computação). 2. Linguagem abusiva. 3. Análise de sentimentos. 4. Preconceito. 5. Mídias. I. Título.

CDU: 004.738.5

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Sebastião Rogério da Silva Neto, intitulada: "Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos", apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 29 de setembro de 2017, às 16h, no Auditório, do Instituto de Computação da UFAL.

COMISSÃO JULGADORA

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Orientador

Prof. Dr. Rafael Ferreira Leite de Mello
Universidade Federal Rural de Pernambuco
Orientador

Prof. Dr. Baldoino Fonseca dos Santos Neto
UFAL – Instituto de Computação
Examinador

Prof. Dr. Rinaldo José de Lima
Universidade Federal Rural de Pernambuco
Examinador

Resumo

Há um interesse crescente na detecção de linguagem abusiva, discurso de ódio, bullying cibernético nos últimos anos. Os sites e redes sociais também sofreram uma pressão cada vez maior para enfrentar esses problemas. O discurso de ódio geralmente é definido como qualquer comunicação que despreza uma pessoa ou um grupo com base em algumas características, como raça, cor, etinidade, gênero, orientação sexual, nacionalidade, religião ou outra característica. Na área de Inteligência Artificial, a mineração de texto pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais. Dentre as técnicas de mineração de texto a Análise de sentimento, ou como também conhecida de Mineração de opinião, atuam com o estudo de opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas. Este trabalho propõe uma possível solução para descoberta de indícios de preconceito em textos em português brasileiro, onde foi desenvolvido uma abordagem híbrida combinando abordagens baseadas em aprendizagem de máquina e dicionários léxicos. Além disso, a abordagem foi utilizada num estudo piloto para identificação de comentários preconceituosos em redações.

Abstract

There is a growing interest in detecting abusive language, hate speech, cyber bullying in recent years. Web sites and social networks have also been under increasing pressure to address these issues. Hate speech is usually defined as any communication that disparages a person or group based on some characteristics, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic. In the area of Artificial Intelligence, text mining can be defined as a set of techniques and processes for discovering innovative knowledge from textual data. Among the text mining techniques is Sentiment Analysis, or also known as opinion mining, act with the study of opinions, feelings, evaluations, attitudes and emotions of people in relation to entities such as products, services, organizations, individuals, problems. This paper proposes a possible solution for the discovery of evidence of prejudice in texts in Brazilian Portuguese, where a hybrid approach was developed combining approaches based on machine learning and lexical dictionaries. In addition, the approach was used in a pilot study to identify biased comments in essays.

Agradecimentos

Quero agradecer a Deus que permitiu e me conduziu a alcançar mais esse sonho.

Dedico essa dissertação a minha família que sempre me apoiou e me incentivou durante todo esse processo.

A minha mãe Noemia Rogério, pelo exemplo de vida e dedicação a educação, que me influenciou a seguir a carreira acadêmica, como também, pela motivação, companheirismo e conselhos durante toda minha vida.

A meu pai Jurandir Severino, por me introduzir ao mundo da informática e sempre me motivar a desbravar essa área.

A meu Tio Sebastião Rogério de Freitas Silva, meu amigo, parceiro, que sempre esteve me ajudando, motivando durante todas as conquistas da minha vida.

A meus orientadores Prof. Evandro e Prof. Rafael que me ajudaram a crescer como pesquisador, pelas contribuições, discussões, lições, meu respeito e apreço.

Ao grupo de pesquisa Tecnologias Inteligentes, Personalizadas e Sociais - TIPS e seus integrantes, em especial meus amigos Tarsis Marinho e Michel Miranda que contribuiram de diversas maneiras na elaboração dessa dissertação como também nas implementações da abordagem.

Não poderia deixar de citar o Laboratório de Excelência em Inteligência Computacional - LEXICO da UFRPE, através de Anderson Pinheiro e Máverick André, que contribuíram nas discussões e implementações da proposta

Ao Programa de Pós-Graduação em Informática - PPGI/UFAL em especial a minha amiga Floripes Teixeira e Vitor Torres. Por fim, a todos os meus amigos que de forma direta ou indireta me ajudaram a realização desse sonho.

Em especial dedico este trabalho a minha Vovó Alaide Cordeiro, que durante esse processo nos deixou, mas seu exemplo de vida e dedicação, serviu como referencial para minha vida, sempre esteve presente em todos os meus avanços acadêmicos, desde o ensino fundamental, médio e superior, agradeço a Deus por todos os momentos desfrutados a seu lado.

"O que as suas mãos tiverem que fazer, que o façam com toda a sua força, pois na sepultura, para onde você vai, não há atividade nem planejamento, não há conhecimento nem sabedoria. Ec 9.10".

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

Há um interesse crescente na detecção de linguagem abusiva, discurso de ódio, bullying cibernético nos últimos anos [39]. Os sites e redes sociais também sofreram uma pressão cada vez maior para enfrentar esses problemas. O discurso de ódio é comumente definido como qualquer comunicação que despreza uma pessoa ou um grupo com base em algumas características, como raça, cor, etinidade, gênero, orientação sexual, nacionalidade, religião ou outra característica [30].

As semelhanças entre as subtarefas levaram os estudiosos a agrupá-los sob os termos de "linguagem abusiva", "discurso prejudicial" e "discurso de ódio" [29], [39], existem trabalhos que já vem sendo desenvolvidos com essas tarefas como por exemplo, Van Hee et al. [45] identifica observações discriminativas (racistas, sexistas) como um subconjunto de "insultos" enquanto Nobata et al. [29] classifica observações semelhantes como "discurso de ódio" ou "linguagem degoratória". Waseem e Hovy [47] consideram apenas "discurso de ódio" sem considerar qualquer potencial sobreposição com bullying ou linguagem ofensiva. A falta de consenso resultou em diretrizes de anotação contraditórias.

Para ajudar a reunir essas literaturas e evitar essas contradições, foi proposto por Wassem et al., [47] uma tipologia que sintetiza essas diferentes subtarefas. As diferenças entre subtarefas em linguagem abusiva podem ser reduzidas a dois fatores principais:

1. O idioma é direto para um indivíduo ou entidade específica ou é direcionado para um grupo generalizado?

2. O conteúdo abusivo é explícito ou implícito?

Cada uma das diferentes subtarefas relacionadas ao idioma abusivo ocupa um ou mais segmentos desta tipologia. O objetivo é esclarecer as semelhanças e diferenças entre subtarefas na detecção de linguagem abusiva para ajudar os pesquisadores a selecionar estratégias adequadas para a anotação e modelagem de dados.

Para combater o idioma abusivo, muitas empresas de internet têm padrões e diretrizes que os usuários devem aderir e empregar editores humanos, em conjunto com sistemas que usam expressões regulares e listas negras, para pegar identificar texto abusivo e assim remover uma publicação.

Definições típicas de discurso de ódio fazem referência ao conteúdo da fala, tom de discurso, avaliações da natureza desse tipo de discurso são também possíveis consequências ou implicações do ato de fala. Para que isso se assenta bem com o domínio da análise do sentimento prevê um modelo de design que possa capturar o conteúdo e aspectos avaliativos do discurso de ódio e desenvolver um sistema que pode ajudar a determinar a gravidade das mensagens de ódio [17].

Na área de Inteligência Artificial, a mineração de texto pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais [37]. Dentre as técnicas de mineração de texto a Análise de sentimento, ou como também conhecida de Mineração de opinião, atuam com o estudo de opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas [22].

A análise do sentimento é a tarefa de identificar opiniões positivas e negativas, emoções e avaliações [18]. A análise do sentimento tem sido usada com muito sucesso nos campos onde os usuários têm uma agenda subjetiva óbvia, como avaliações de filmes [14] ou blogs [13]. É muito menos claro como as técnicas de análise do sentimento podem ser empregadas no contexto da análise de rede social, onde a linguagem tende a ser mais livre e informal [10].

O discurso de ódio e a análise do sentimento estão intimamente relacionadas, e é seguro assumir que geralmente o sentimento negativo pertence a uma mensagem de discurso de ódio. Devido a isso, várias abordagens reconhecem a relação de discurso de ódio e análise de sentimentos, incorporando o último como uma classificação auxiliar [14] e [17] seguem uma abordagem de vários passos, em que um classificador dedicado a detectar polaridade negativa

é aplicado antes do classificador, verificando especificamente a evidência de discurso de ódio.

As opiniões são fundamentais para quase todas as atividades humanas, pois são influenciadores de comportamentos. Sempre que é preciso tomar uma decisão, busca-se conhecer outras opiniões. Empresas e organizações sempre querem encontrar opiniões de consumidores ou públicos sobre seus produtos e serviços. Os consumidores individuais também querem conhecer as opiniões sobre candidatos políticos por exemplo, antes de tomar uma decisão de voto em uma eleição política. Há pouco tempo, quando um indivíduo precisava de opiniões, era necessário procurar amigos e familiares. Quando uma organização ou uma empresa precisava de opiniões públicas ou de consumidores, realizava pesquisas de opinião e grupos focais [22].

Diversas aplicações e pesquisas tem sido desenvolvidas nesse contexto, Liu et al., [23] propos um modelo de sentimento para prever o desempenho das vendas. McGlohon, Glance e Reiter, [25], utilizaram revisões de textos para classificar produtos e comerciantes. O'Connor et al., [31] e Tumasjan et al., [44] observaram que os sentimentos de *tweets* estavam vinculados às pesquisas de opinião pública e a resultados eleitorais, respectivamente. Também, um método para prever volumes de comentários de blogs políticos foi relatado por Yano e Smith, [50]. Finalmente, dados do Twitter, críticas de filmes e blogs foram usados para prever receitas de bilheteria para filmes Asur e Huberman, [5]; Joshi et al., [21]; Sadikov, Parameswaran e Venetis, [38].

1.1 Discurso de ódio

Nas mídias sociais, o discurso de ódio é uma espécie de escrita que deprecia e é sucetível a causar danos ou perigo à vítima. É um viés hostil, discurso malicioso dirigido a uma pessoa ou a um grupo de pessoas devido algumas características reais ou inatas. É um tipo de discurso que demonstra uma clara intenção de ser prejudicial, incita danos ou promove o ódio. O ambiente das mídias sociais fornece um terreno particularmente fértil para criação, compartilhamento e troca de mensagens de ódio contra um grupo inimigo percebido. Esses sentimentos são expressos em sites de revisão de notícias, fóruns na internet, discussão grupos, bem como em sites de microblog.

De acordo com [17] o discurso de ódio pode ser identificado da seguinte maneira:

(1) É direcionado a um grupo de pessoas e não a uma pessoa. Discurso de ódio ou perigoso é um discurso prejudicial que chama o público a tolerar ou participar de atos violentos contra um grupo de pessoas. Assim, no espaço online os tipos mais comuns de ódio, o discurso está relacionado à nacionalidade, etnia, religião, gênero, orientação social, deficiência e classe.

(2) Esse tipo de discurso pode conter algumas das características, como declarações que comparam um grupo de pessoas com vermes ou insetos (metafóricos), sugerem que a audiência enfrenta uma séria ameaça ou violência de outro grupo e sugere que algumas pessoas de outro grupo estragam a pureza ou a integridade dos autores do grupo.

(3) O discurso perigoso geralmente encoraja o público a adotar ou a cometer atos violentos no grupo visado. Os seis chamados como as ações comuns de discurso de ódio são: discriminar, saquear, revirar, vencer, expulsar com força e matar.

Nesse contexto a utilização de técnicas de análise de sentimento no contexto de linguagem abusiva podem ser combinadas para identificação de indício de preconceito que ainda é uma linha que foi pouco explorada ao longo dos anos no idioma português brasileiro. Como já comentado anteriormente na internet é cada vezes mais comum que os usuários expressem suas opiniões e em alguns casos, com cunho preconceituoso deixando as redes sociais, blogs, fóruns vulneráveis a ataques individuais ou até mesmo de grupo radicais.

Existem trabalhos na literatura que apresentaram resultados interessantes sobre a identificação de preconceito, linguagem abusiva, porém, todos eles encontram-se na maior parte no idioma inglês, para língua portuguesa e principalmente no contexto educacional há pouquíssimos trabalhos, sendo assim, esse trabalho propõe uma possível solução para descoberta de indícios de preconceito em textos em português brasileiro, para isso, foi conduzido um estudo de caso no Exame Nacional do Ensino Médio (ENEM), especificamente nas redações, considerando a competência 5 que será descrita no estudo, no que diz respeito aos direitos humanos.

1.2 Problemática

Como foi destacado na seção anterior, técnicas de análise de sentimento podem ajudar a detectar comentários abusivos, preconceituosos em texto.

Detectar linguagem abusiva é muitas vezes mais difícil do que se espera por uma variedade de razões. O barulho nos dados em conjunto com a necessidade de conhecimento não só torna essa tarefa desafiadora para automatizar, mas também potencialmente uma tarefa difícil para as pessoas também.

Afim de identificar opiniões são utilizadas abordagens baseadas em aprendizagem de máquina ou técnicas que utilizam dicionários lexicos [24] cada termo lexico é identificado como um lexema, que consiste de uma forma ortográfica e fonológica com uma forma de representação de significado.

Para a extração de opinião normalmente subjetivos em textos, é feito a sumarização das opiniões realizada através das classificações das opiniões com base em categorias (Polaridade): positiva, negativa e neutra [27].

Um exemplo de dicionário léxico é o WordNet Affect BR que consiste em uma base com palavras de emoções na língua portuguesa [34]. Nessa base são apresentados grupos de palavras relacionadas a diferentes emoções.

A Análise de sentimento pode ser utilizada para identificar preconceito. Nela as técnicas são aplicadas para identificar textos positivos (sem preconceito, normalmente chamado de texto neutro) ou negativo (texto com comentários preconceituosos). Existem dicionários léxicos específicos para termos preconceituosos [17; 43].

Ao observar a literatura, podemos destacar alguns problemas, que tem-se destacado, entre eles estão:

- **Preconceito explícito e implícito:** normalmente o texto pode ser classificado como explícito, quando há de forma clara e objetiva termos abusivos, nos casos em que é apresentado de forma implícita, o usuário geralmente é sarcástico, duvidoso no texto, dificultando a identificação do indício dos termos abusivos.

- **Subjetividade no texto:** é um problema bastante interessante, pois na maioria dos casos o texto é opinativo, tornando-o subjetivo e de difícil identificação.

- **Termos preconceituosos em português:** na maioria das pesquisas citadas, é apresentado dicionários com termos preconceituosos mas na língua inglesa, existe uma lacuna evidente, no que diz respeito a um dicionário especificamente na língua portuguesa.

- **Difícil rastreamento dos insultos raciais e minoritários**: Desenvolver um classificador de abuso ou palavrões razoavelmente eficaz com uma lista negra (uma coleção de palavras conhecidas por possuirem ódio ou insultos), no entanto, essas listas não são estáticas e estão mudando. Portanto, uma lista negra deveria ser atualizada regularmente para acompanhar a mudança de idioma. Além disso, alguns insultos que podem ser inaceitáveis para um grupo podem ser totalmente aceitável para outro grupo, e, portanto, no contexto da lista negra tudo é importante.

- **Sarcasmo**: Finalmente, observamos casos em que alguns usuários publicariam comentários sarcásticos na mesma voz que as pessoas que estavam produzindo linguagem abusiva. Isso é muito difícil para os seres humanos ou as máquinas corrigirem, pois requer conhecimento da comunidade e até mesmo dos próprios usuários.

1.3 Questão de Pesquisa

A pergunta da pesquisa que vai direcionar o trabalho é: Como identificar termos preconceituosos em textos em português que venham a ferir/transgredir os direitos humanos?

1.4 Objetivos

1.4.1 Geral

Desenvolver uma abordagem computacional para identificação de indícios de preconceito na língua portuguesa baseado em aprendizado de máquina e dicionários léxicos.

1.4.2 Específicos

- Identificar principais grupos de palavras com emoções com maior índice de preconceito.

- Criação de um dicionário para identificação de preconceito em português.
- Análise de diferentes algoritmos para identificação de frases com indício de preconceito em redações em português
- Combinar técnicas de aprendizagem de máquina com dicionários léxicos para identificação de preconceito.

1.5 Relevância

Com o crescimento constante de pesquisas em linguagem abusiva, com tarefas de:

- Identificação de dicurso de ódio
- Preconceito explícito e implícito

O aumento maciço dos conteúdos web gerados pelos usuários, em particular nas redes de redes sociais, a quantidade de discurso de ódio também está aumentando constantemente. Ao longo dos últimos anos, o interesse pela detecção de fala de ódio, em particular, a automatização desta tarefa tem crescido continuamente, juntamente com o impacto social deste fenômeno [39].

Tendo em vista a grande quantidade de informações armazenadas em formato textual não estruturado, a Mineração de Texto através das técnicas de Análise de sentimentos apresenta-se como um campo de pesquisas importante, sendo útil e aplicável a diversas áreas.

Motivado por essas circunstâncias, é apresentado neste trabalho o desenvolvimento de uma abordagem híbirida que utiliza algoritmos de aprendizagem de máquina em conjunto com dicionários léxicos com termos abusivos em português para identificação de indícios de preconceito em textos na língua portuguesa.

1.6 Organização do texto

Essa dissertação está dividida em seis capítulos. O capítulo 1 introduz a contextualização, problemática e os objetivos do trabalho proposto. No capítulo 2 são apresentados o referencial teórico com conceitos relacionados ao tema deste trabalho. A conceitualização de

mineração de texto, técnicas de classificação utilizadas no trabalho, mineração de opinião e identificação de preconceito. No mesmo capítulo a subseção 2.3.2 em diante definem a abordagem baseada em aprendizagem de máquina e em dicionário, concluindo com a abordagem híbrida onde é apresentado a combinação entre as abordagens.

O capítulo 3 apresenta os trabalhos relacionados que utilizam abordagens baseadas em aprendizagem de máquina, como também, com dicionários, foi feita uma descrição da metodologia dos trabalhos. No fim do capítulo uma tabela apresenta as principais diferenças do trabalho com as principais abordagens encontradas.

No capítulo 4 definido de metodologia é apresentado as etapas para o desenvolvimento do trabalho proposto, dividindo-se em quatro etapas: coleta da base de dados, pré-processamento, identificação de termos preconceituosos, classificação. O capítulo 5 expõe o estudo de caso conduzido na proposta.

No capítulo 6 apresenta as considerações finais, limitações e os trabalhos futuros.

Capítulo 2

Referencial Teórico

Este capítulo tem por objetivo apresentar os principais tópicos que foram a base teórica para o entendimento do trabalho aqui proposto.

2.1 Mineração de Texto

A mineração de textos é um ramo da mineração de dados [11], pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais [37].

Segundo [28] o processo de mineração de textos pode ser dividido em várias etapas, dentre elas há a extração de padrões válidos, onde são descobertos padrões válidos e úteis nos textos. Essa etapa pode ser dividida em classificação, sumarização de textos e agrupamento.

- Classificação

Para [8] a classificação ou categorização é um processo que visa a identificação de tópicos principais em um documento e a sua associação baseando-se em um algoritmo pré-definido, construído a partir de um conjunto de treinamento definido por pessoas experientes no assunto envolvido.

- Sumarização

Na sumarização é feito a seleção das informações mais importantes do texto, tornando a descrição mais compacta, porém mantendo a mesma informação [8].

- Agrupamento

Para [37] em tarefas de agrupamento, um conjunto de documentos são disponibilizados e são divididos em grupos, onde documentos do mesmo grupo são altamente similares entre si, mas dissimilares em relação aos documentos de outros grupos.

2.2 Técnicas de Classificação

Nesta seção são apresentados resumidamente os classificadores adotados para realização dos experimentos.

2.2.1 Redes Bayesianas

Esse classificador é uma técnica probabilística baseada no Teorema de Bayes [3]. Ele calcula a probabilidade que uma amostra desconhecida pertença a cada uma das classes possíveis, predizendo a classe mais provavél. Para isto, o classificador baseado em rede baysesiana calcula uma distribuição geradora para cada classe do problema através da análise das relações entre as características envolvidas e as classes de cada instância.

Para [16] ele funciona da seguinte maneira: primeiro, em uma fase de aprendizado, ele gera uma lista de palavras com suas frequências a partir do corpus de entrada. Cada palavra nesta lista gerada é rotulada com a classe a que pertence, resultando em uma espécie de dicionário para cada classe. A partir deste dicionário de classes, é construída uma árvore cujas folhas são classes e nós intermediários que indicam probabilidades calculadas de acordo com o teorema de Bayes. Por fim, quando um novo texto é apresentado ao classificador, ele percorre a árvore gerada até encontrar uma folha que indica a classe que mais combinam com as palavras no texto de entrada.

2.2.2 Máquina de Vetor de Suporte

A máquina de vetor de suporte (SVM) é um classificador baseado na teoria de aprendizado estatístico de Vapnik [46]. Para efetuar classificações/reconhecimento de padrões o SVM constrói hiperplanos em um espaço multidimensional objetivando separar casos de diferentes classes. Cada hiperplano é considerado como uma separação ótima que separa os vetores

das classes sem erro e com distância máxima para com os vetores mais próximos. É considerado como um método espaço-vetorial para problemas de classificação binária. Segundo [6] os termos de índice compõem um espaço *t*-dimensional no qual os documentos são representados como pontos (ou vetores). Dadas as representações vetoriais para os documentos, a ideia é encontrar uma superfície de decisão (um hiperplano) que pode ser usado para melhor separar os elementos em duas classes *ca* e *cb*.

O hiperplano, o qual é aprendido a partir de dados de treinamento, divide o espaço em duas regiões, tal que todos os documentos na classe *ca* estejam em uma região e todos os documentos na classe *cb* estejam na outra região. Em um espaço de duas dimensões, esse hiperplano é uma linha. Em um espaço tridimensional, esse hiperplano é um plano. Uma vez que o hiperplano tenha sido aprendido, um novo documento *dj* pode ser classificado pela sua posição relativa ao hiperplano [6].

Para ilustrar segue a Figura 2.1, abaixo com um simples bidimensional cujos pontos dos dados de treinamento sejam linearmente separáveis. Nesse exemplo bidimensional, a linha *s* maximiza as distâncias aos documentos mais próximos nas classes *ca* e *cb* (compare com a linha r) e constituem o hiperplano de decisão usado para classificar novos documentos.

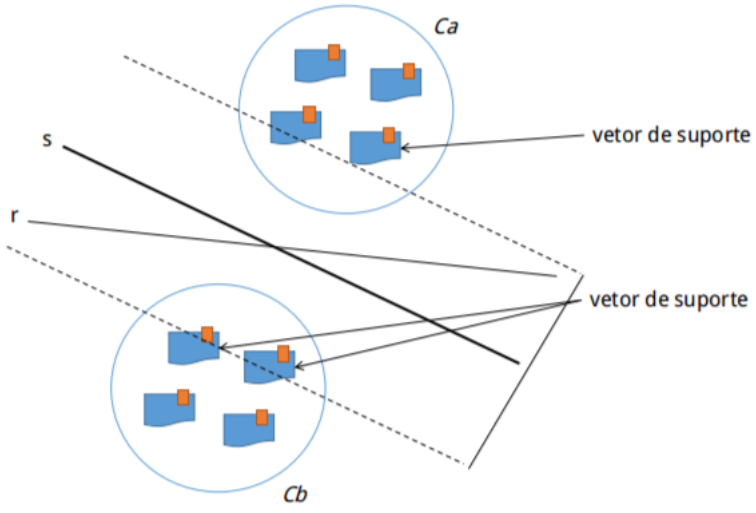

Figura 2.1: Hiperplano de decisão Baeza-Yates, R., Ribeiro-Neto, B. (2013)

As linhas paralelas pontilhadas delimitam a região onde devemos procurar por uma solução. Por conveniência, nos referimos a eles como *hiperplanos delimitadores*. Um documento que pertence a um hiperplano delimitador é chamado de *vetor de suporte*. Linhas paralelas

aos hiperplanos delimitadores são as melhores candidatas. No exemplo, a linha *s* que separa o espaço em porções iguais é o melhor hiperplano, ela é o *hiperplano de decisão*

2.2.3 Redes Neurais Artificiais

Rede neural artifical é uma técnica de aprendizagem de máquina que simula o funcionamento de um sistema nervoso [3]. Para isso, conta com a presença de neurônios artificiais interligados entre si por meio de sinapses (na computação pesos). Cada neurônio recebe entradas e, associados a estas, pesos que representam a força do sinal sináptico. A partir das entradas e de seus respectivos pesos, um somatório ponderado é realizado no núcleo do neurônio e com base em um limiar de ativação é verificado se a entrada será ou não propagada para neurônios das camadas adjacentes a camada atual.

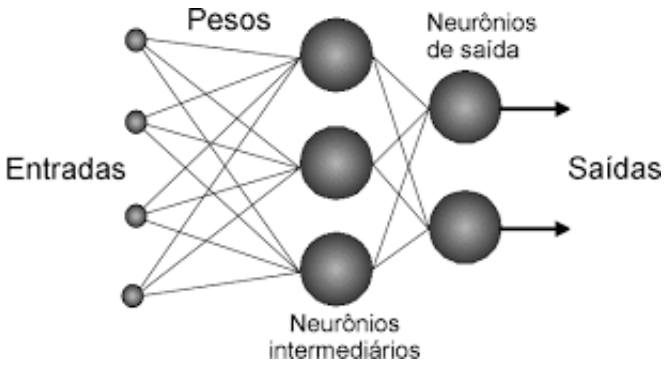

Figura 2.2: Exemplo de rede neural artifical de 2 camadas com 4 entradas e 2 saídas Tafner, M. A. (1998).

2.2.4 K-Nearest Neighbors - KNN

O *K-Nearest Neighbors* técnica de classificação de padrões que consiste em atribuir uma classe a um elemento desconhecido usando a classe da maioria de seus vizinhos mais próximos, segundo uma determinada distância (no espaço de atributos). O algoritmo KNN é baseado em analogia, um objeto é classificado pelo voto da maioria de seus vizinhos [42].

Segundo [16] o classificador funciona da seguinte maneira: primeiro, é gerado uma lista de palavras com suas frequências a partir do *corpus*, com estes dados, um vetor é gerado para cada documento. Então, esses vetores são colocados em um plano cartesiano criado pelo

algortimo. Quando um novo documento é apresentado ao classificador, que gera um vetor que representa este novo documento e a distância calculada para todos os vetores do plano. A classe com mais votos ao longo dos k vizinhos mais próximos será a classe escolhida para o documento de entrada.

2.2.5 Árvore de Decisão

É uma técnica de aprendizado de máquina que utiliza uma estrutura de árvore para avaliar os atributos de uma entrada e retorna uma predição baseada nos valores desses atributos. A árvore é estruturada através de vários nós, onde cada nó corresponde a um teste do valor de uma característica do dado de entrada. Os nós da árvore são ligados por ramos, os quais identificam os possíveis valores do teste realizado em cada nó. Por fim, cada nó da folha da árvore representa um valor de retorno [3].

Árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema (Gama, 2004).

2.3 Conjuntos de características

2.3.1 Mineração de Opinião e Identificação de Preconceito

A mineração de opinião pode ser caracterizada a partir de três tarefas [9]:

a) identificar (tópicos, sentenças opinativas): que consiste em encontrar os tópicos existentes, e possivelmente associá-los com o respectivo conteúdo subjetivo.

b) classificar a polaridade do sentimento: que classifica um dado texto em uma de duas classes: positivo ou negativo. No entanto, classes adicionais podem ser consideradas para que a análise seja mais robusta, ou para aumentar o nível de detalhe dos resultados. Assim, estas classes podem ser desdobradas em classificações com diferentes graus de intensidade (muitoPositivo, moderadamentePositivo), ou em intervalos numéricos representando um grau de intensidade.

c) sumarizar, para identificação de opinião média ou prevalecente de um grupo de

pessoas sobre um determinado tópico/entidade: a opinião expressa por uma única pessoa não é suficiente, sendo necessário analisar uma grande quantidade de opiniões. É necessário a criação de métricas e sumários que quantifiquem a diversidade de opiniões encontradas a respeito um mesmo alvo. Este é o objetivo desta etapa, onde são criadas métricas que representem o sentimento geral, as quais podem ser visualizadas ou servir de entrada para outras aplicações. A figura 2.2, apresenta essas tarefas.

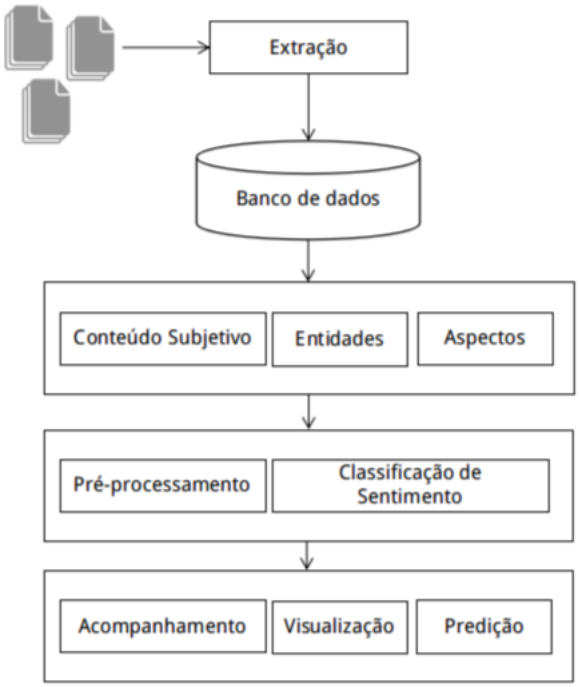

Figura 2.3: Etapas da mineração de opinião

Para a extração de opinião normalmente subjetivos em textos, é feito a sumarização das opiniões realizada através das classificações das opiniões com base em categorias (Polaridade): positiva, negativa e neutra [27].

Afim de identificar opiniões são utilizadas abordagens baseadas em aprendizagem de máquina ou técnicas que utilizam dicionários lexicos [24]. Nele cada termo lexico é identificado como um lexema, que consiste de uma forma ortográfica e fonológica com uma forma de representação de significado.

2.3.2 Recursos Máximos por Documento - MFD

A estratégia de Recursos Máximos por Documento (MFD) aumenta o número de recursos selecionados por documento, com o objetivo de melhorar a precisão do processo de agrupamento. MFD calcula as pontuações de cada recurso no conjunto de recursos e, em seguida, seleciona os recursos f com as pontuações mais altas para cada documento no conjunto de coleções para compor o conjunto final de recursos [35].

O Algoritmo 1, apresenta o pseudocódigo para este método. Ele funciona da seguinte maneira, seleciona apenas um recurso, que extrai os melhores recursos classificados de cada documento *di* [35].

Algoritmo 1: MFD

Require: n {number o feature per documental}
1: load all documents from dataset D_n
2: **for** $W_h \epsilon V$ **do**
3: S_h = *ranking* (W_h)
4: **end for**
5: **for all** $d_i \epsilon D_n$ **do**
6: *topfeature* = $\int FeaturesRank(d_i, \int)$
7: *FS* = *add(topfeature)*
7: **end for**

2.3.3 Características de N-gram

Foi empregado como característica n-gramas (de 4 a 5 caracteres, com espaços incluídos) e unigrams e bigrams. Foi ignorado o texto não normalizado foi utilizado medidas de distância de distância simples para normalizá-los, usamos n-gramas de caracteres para modelar os tipos de abusos conscientes ou inconscientes de palavras ofensivas baseados na proposta de [29]

2.3.4 Características Linguísticas

Afim de lidar ainda mais com o barulho dos dados, foi utilizado características especializadas com base no trabalho de [29]. Essas características visam procurar explicitamente palavras abusivas (como o uso de listas de ódio pré-existentes), mas também elementos de linguagem não abusiva, como o uso de palavras de cortesia ou verbos modais.

2.4 Análise de sentimento

A tarefa de analisar opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas é conhecida por mineração de opinião ou análise de sentimento [22].

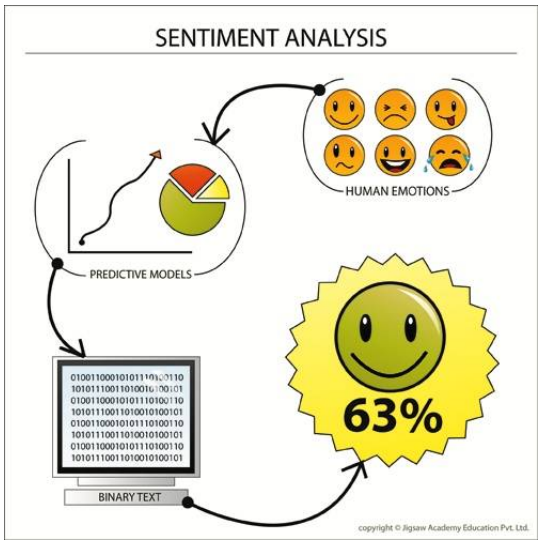

Figura 2.4: Processo da análise de sentimento

2.4.1 Subjetividade e emoção

Existem dois conceitos importantes que estão relacionados com o sentimento e opinião, que são, subjetividade e emoção.

Para [22] uma sentença objetiva apresenta algumas informações reais sobre o mundo, enquanto uma sentença subjetiva expressa alguns sentimentos, opiniões ou crenças pessoais.

Um exemplo de frase objetiva é: “O iPhone é um produto da Apple”. Um exemplo de frase subjetiva é: “Eu gosto do iPhone”. As expressões subjetivas vêm em muitos formas,

por exemplo, opiniões, alegações, desejos, crenças, suspeitas e especulações. Existe uma certa confusão entre os pesquisadores para equiparar a subjetividade com a opinião. Por opinião, é definido uma frase expressa ou implica em um sentimento positivo ou negativo. Os dois conceitos não são equivalentes, embora tenham uma grande interseção. A tarefa de determinar se uma frase é subjetiva ou objetiva é chamada classificação de subjetividade [22]. Aqui, devemos observar o seguinte:

- Uma sentença subjetiva pode não expressar qualquer sentimento. Por exemplo: "eu acho que ele foi para casa" é uma frase subjetiva, mas não expressa nenhum sentimento.

- Frases objetivas podem implicar opiniões ou sentimentos devido a fatos desejáveis e indesejáveis.

As emoções são sentimentos e pensamentos subjetivos. A emoção é estudada em múltiplos campos, por exemplo, psicologia, filosofia e sociologia. Os estudos são muito amplos, a partir de respostas emocionais de reações fisiológicas (por exemplo, mudanças de freqüência cardíaca, pressão sanguínea, suor e assim por diante), expressões faciais, gestos e posturas para diferentes tipos de experiências subjetivas do estado de espírito de um indivíduo. Os cientistas classificaram as emoções das pessoas em algumas categorias. No entanto, ainda não existe um conjunto de emoções básicas acordadas entre os pesquisadores [22].

Com base em [33], as pessoas têm seis emoções primárias, ou seja, amor, alegria, surpresa, raiva, tristeza e medo, que podem ser subdivididos em muitas emoções secundárias e terciárias. Cada emoção também pode ter diferentes intensidades.

As emoções estão intimamente relacionadas aos sentimentos. A força de um sentimento ou opinião é tipicamente ligada à intensidade de certas emoções, por exemplo, alegria e raiva. As opiniões que estudadas na análise do sentimento são principalmente avaliações (embora nem sempre). De acordo com a pesquisa do comportamento do consumidor, as avaliações podem ser amplamente categorizadas em dois tipos: avaliações racionais e avaliações emocionais [12].

Avaliação racional: tais avaliações são de raciocínio racional, crenças tangíveis e atitudes utilitárias. Por exemplo, as seguintes frases expressam avaliações racionais: "A voz deste telefone é clara", "Este carro vale o preço" e "Estou feliz com este carro".

Avaliação emocional: tais avaliações são de respostas não tangíveis e emocionais a entidades que se aproximam do estado de espírito das pessoas. Por exemplo, as seguintes frases expressam avaliações emocionais: “Eu adoro o iPhone”, “Estou tão bravo com o serviço de pessoas” e “Este é o melhor carro já construído”.

2.4.2 Abordagem baseada em aprendizagem de máquina

As técnicas de aprendizado de máquina tem por objetivo descobrir automaticamente regras gerais em grandes conjuntos de dados, que permitam extrair informações implicitamente representadas. De modo geral, essas técnicas podem ser divididas em dois tipos: aprendizado supervisionado e aprendizado não supervisionado [41].

Na área de mineração de opiniões, há um predomínio do uso de métodos supervisionados de aprendizagem, mais especificamente, classificação e regressão. Neste contexto, o problema de classificação é dividido em dois passos: (1) aprender um modelo de classificação sobre um corpus de treinamento previamente rotulado com as classes consideradas (positivo, negativo); e (2) prever a polaridade de novas porções de texto com base no modelo resultante [9].

A qualidade do modelo preditivo resultante da etapa de aprendizagem é medida em termos de métricas como acurácia (capacidade do modelo de prever corretamente), precisão (número de instâncias previstas corretamente em uma dada classe), ou revocação (número de instâncias de uma dada classe previstas na classe correta).

Uma das grandes limitações no uso de aprendizado supervisionado para definição de polaridade é a necessidade de dados rotulados para treino. O desempenho destes métodos é afetado não somente pela quantidade, mas igualmente pela qualidade dos dados de treino disponíveis. Ainda, cada conjunto de treino é fortemente vinculado ao seu domínio [9].

2.4.3 Abordagem baseada em dicionário

A abordagem baseada em dicionário também conhecida como léxica ou linguísitca, apresenta o uso de léxicos (dicionários) de sentimentos, que são compilações de palavras ou expressões de sentimento associadas à respectiva polaridade [9].

É comum nesta abordagem a utilização do método de co-ocorrência entre alvo e senti-

mento, que não leva em consideração nem a ordem dos termos dentro de um documento, nem suas relações léxico-sintáticas. A classificação de sentimento em texto é simples, caso exista uma palavra de sentimento, com sua polaridade dada por um léxico de sentimentos. Esse método é extensamente empregado para o atrelamento de um sentimento a uma entidade em uma sentença [9].

O método por co-ocorrência apresenta bons resultados quando o nível de análise textual é de granularidade pequena, pois a palavra detentora do sentimento está próxima à entidade que qualifica. Sendo assim, este método é usualmente utilizado em análises de nível de sentença, cláusula ou até em documentos com poucos caracteres, como um *tweet* conforme [9].

Para um nível maior de granularidade, é estabelico uma média sobre as palavras de sentimentos encontradas. Abaixo é apresentado a equação com uma função genérica de determinação de polaridade em um documento D, onde *Sw* representa a polaridade de uma palavra *w* em um dicionário. A agregação pode levar em conta funções de peso e de modificação. A função *peso()* pode ser, por exemplo, alguma medida de distância entre a palavra de sentimento e o alvo, ou de importância da palavra no texto (frequência). A função *modificador()* pode ser usada para tratar negações, palavras de intensidade, etc. Esta função de agregação também pode ser estendida a sentenças, cujas cláusulas podem combinar diferentes palavras de sentimento.

$$\mathbf{S}(\mathbf{D}) = \frac{\Sigma w \epsilon D Sw \cdot peso(w) \cdot modificador(w)}{\Sigma peso(w)}$$

Outros métodos linguísticos mais complexos, como a utilização de *parsers* linguísticos, que têm como propósito analisar o texto e aumentar a qualidade da classificação com base em informações morfossintáticas ali presentes (sujeito, predicado, dependências, funções sintáticas, etc.). No entanto, ferramentas de processamento de linguagem natural são em sua maioria restritas a determinado idioma. Recursos para a língua portuguesa são escassos, quando comparada à língua inglesa, situação esta comum a outra línguas [9].

2.4.4 Abordagem híbrida

A abordagem híbrida considerada neste trabalho, consiste na combinação da abordagem baseada em aprendizagem de máquina com a abordagem baseada em dicionários. Existem

trabalhos como já discutido anteriormente que utilizam apenas uma das abordagens, a combinação delas até então, foi pouco explorada na literatura.

A combinação consiste em utilizar dicionários (léxicos), que são compostos de palavras ou expressões de sentimentos positivos, negativos ou neutros já utilizados na literatura delimitando a termos preconceituosos, com esses léxicos delimitados a aprendizagem de máquina através das técnicas de classificação pode ser utilizada para identificação de indícios de preconceito no texto.

Capítulo 3

Trabalhos Relacionados

Este capítulo tem como objetivo apresentar uma comparação entre abordagens similares a este trabalho, bem como os diferenciais desta proposta. São apresentados os detalhes das soluções propostas, inicialmente, serão apresentadas propostas que utilizam análise de sentimento, por conseguinte, trabalhos que tratam em específico identificação de preconceito. Por fim, é sintetizada na Tabela 1, as características definidas como essenciais para atender os desafios supracitados, em relação aos trabalhos relacionados e a abordagem ora proposta.

Um dos primeiros trabalhos encontrados na literatura propõe a utilização das palavras do texto como entrada para um classificador identificar o comentário racista [18]. Em outro trabalho mais recente o classificador utiliza diferentes características do texto como: tópicos, sentimento e marcações semânticas [1].

Outros trabalhos utilizam dicionários léxicos para realizar a análise, ambos tratam principalmente de preconceito relacionados a raça, nacionalidade e religião [17; 43]. Gitari *et al.* [17] propõe a extensão de dicionários léxicos como WordNet e SentiWordNet para a criação de um novo dicionário focado na identificação de palavras relacionadas a preconceito. Enquanto em [43] é proposto um dicionário de termos preconceituosos em Alemão. Em [10] é combinado técnicas de análise de redes sociais e sentimento para identificar comentários racistas .

Como não foi encontrado trabalhos que tratassem diretamente da identificação de preconceito em textos educacionais, alguns trabalhos considerados relevantes na construção deste capitulo utilizaram técnicas de análise de sentimento e dicionários léxicos.

Coutinho *et al.,* [13] apresenta uma análise de sentimento em mensagens de texto, para

isso foram capturadas mensagens de um chat de um Ambiente Virtual de Aprendizagem - AVA, foi determinado a polaridade das mensagens como positiva, negativa e neutra e classificadas com o algorítmo Naive Bayes que utiliza probabilidades para determinar qual categoria melhor se enquadra a um determinado texto de entrada.

No trabalho de Dosciatti *et al.,* [15] é apresentado uma abordagem utilizando Máquinas de Vetores de Suporte (SVM) na identificação de emoções em textos escritos em Português do Brasil. A base de dados utilizada no experimento foi composta por notícias extraídas de um jornal online. Os textos foram previamente rotulados e submetidos a um classificador SVM com configuração multiclasse, obtendo uma taxa de acerto de 61%.

Agarwal *et al.,* [2; 1] apresenta uma pesquisa onde os usuários abusam da liberdade de expressão para publicar comentários preconceituosos sobre várias religiões e raças. Neste artigo, é definido o problema de radicalização e detecção de racismo como um problema de classificação, baseado em intenção que identifica um *post* para ser radicalizado ou racista com base no motivo do autor. É apresentado a eficácia dos traços de personalidade dos autores para identificar uma postagem com tal intenção. Na seção 3.1 o trabalho será apresentado mais detalhadamente.

Afim de otimizar a mediação pedagógica do professor no acompanhamento das produções textuais, Panceri *et al.,* [32] propõe a construção do núcleo de processamento de textos do SMA Alpes, através da combinação de técnicas de recuperação de informação, mineração de textos, análise semântica latente e clusterização. Para verificar a aderência da proposta foi aplicado o SMA Alpes a um debate de teses realizado em 2013, apresentando resultados satisfatórios.

Em [29] é discutido a detecção de linguagem abusiva no conteúdo online gerado pelo usuário que tem-se tornado uma questão de crescente importância nos últimos anos. A maioria dos métodos comerciais atuais faz uso de listas negras e expressões regulares, no entanto, essas medidas caem quando se disputa com exemplos mais sutis e menos presunçosos de discurso de ódio. Neste trabalho, foi desenvolvido um método baseado em aprendizado de máquina para detectar discurso de ódio em comentários de usuários on-line de dois domínios que superam uma abordagem de aprendizagem profunda de última geração. Também desenvolveram um corpus de comentários de usuários anotados para linguagem abusiva, o primeiro de seu tipo. Por fim foi utilizado, a ferramenta de detecção para analisar o idi-

oma abusivo ao longo do tempo e em diferentes configurações para aprimorar ainda mais o conhecimento desse comportamento.

Todos os dados utilizados para treinar e testar, foram extraídos dos comentários encontrados no Yahoo! (Finanças e Notícias). Esses comentários foram moderados pelos funcionários do Yahoo. Todas as disciplinas possuíam pelo menos um grau de graduação e estavam familiarizadas com o conceito de julgar passagens de texto para diferentes tipos de tarefas e requisitos de anotação. Antes de assumir a moderação real tarefa, eles foram treinados para se familiarizarem com as diretrizes de julgamento de texto.

Neste trabalho, foi empregado um método de classificação supervisionado que utiliza características de PLN que medem diferentes aspectos do comentário do usuário. Especificamente, foi utilizado o modelo de regressão Vowpal Wabbit5 em sua configuração padrão com uma taxa de bits de 28. As características podem ser divididas em quatro classes: N-gramas, Semântica Linguística, Sintatica e Distributiva. Para os três primeiros recursos, foi feito um pré-processamento suave para transformar alguns dos ruídos encontrados nos dados que podem afetar a quantidade de recursos esparsos no modelo. As transformações de exemplo incluem números de normalização, substituindo palavras muito longas desconhecidas pelo mesmo token, substituindo a pontuação repetida com o mesmo token.

O trabalho de [39] apresenta uma pesquisa sobre detecção de fala de ódio. Dado o crescente crescimento do conteúdo de mídias sociais, a quantidade de discurso de ódio em linha também está aumentando. Devido à enorme escala da web, são necessários métodos que detectem automaticamente a fala odiosa. Na pesquisa é descrito as áreas-chave que foram exploradas para reconhecer automaticamente esses tipos de enunciados usando o processamento do idioma natural. como também são discutidos os limites dessas abordagens.

As pessoas ficaram mais ansiosas para expressar e compartilhar suas opiniões na web sobre atividades do dia-a-dia e questões globais também. A evolução das mídias sociais também contribuiram imensamente para essas atividades, proporcionando-nos assim uma plataforma transparente para compartilhar visões em todo o mundo. Essas declarações eletrônicas expressadas na web são muito prevalentes na indústria de negócios e serviços para permitir que o cliente compartilhe seu ponto de vista. Na última década e meia, as comunidades de pesquisa, a academia, as indústrias públicas e de serviços estão trabalhando rigorosamente na análise de sentimentos, também conhecida como mineração de opinião,

para extrair e analisar o clima e os pontos de vista públicos. A este respeito, o trabalho de [36] apresenta uma pesquisa rigorosa sobre a análise do sentimento, que retrata os pontos de vista apresentados por mais de cem artigos publicados na última década sobre as tarefas, abordagens e aplicações necessárias da análise do sentimento. Várias sub-tarefas precisam ser realizadas para a análise do sentimento, que por sua vez pode ser realizada usando várias abordagens e técnicas. Esta pesquisa abrange a literatura publicada durante 2002-2015, é organizada com base em subtarefas a serem executadas, técnicas de processamento de máquina e linguagem natural utilizadas e aplicações de análise de sentimentos. O artigo também apresenta questões abertas e, juntamente com uma tabela de resumo de cento e sessenta e um artigos.

A proposta deste trabalho é desenvolver uma abordagem híbrida que utiliza agoritmos de aprendizagem de máquina e dicionários léxicos para identificação de textos com indício de preconceito. Desta forma, este trabalho se diferencia dos encontrados na literatura, por combinar dicionários léxicos com algoritmos de AM.

Baseado nos útlimos trabalhos citados foram selecionados alguns que mais se assemelhavam com a presente proposta. Nas próximas seções serão descritos estes trabalhos, onde serão discutidos as principais características de cada uma delas.

3.1 But I did not Mean It! - Intent Classification of Racist Post on Tumblr

A pesquisa de Agarwal [1] discute mentalidade semelhante de usuários que usam sites populares para postar discurso preconceituoso contra várias religiões e raças. A identificação automática de postagens racistas e de promoção do preconceito é necessária para a construção de redes sociais mais seguras. No entanto, apenas as técnicas baseadas em palavras-chave não podem ser usadas para identificar com precisão a intenção de uma publicação, pois existe o desafio de determinar a presença de ambiguidade nessas publicações.

Para tanto, neste trabalho foi conduzido um experimento através do microblog Tumblr, onde foi desenvolvido um classificador de aprendizado de máquina para identificar os *posts* com intenção racista. No treinamento do modelo, foi identificado vários sentimentos semânticos, linguísticos a partir de texto livre. Os resultados experimentais mostram que a

abordagem proposta é efetiva, sendo que as tendências sociais, as pistas da linguagem e os traços de personalidade de uma narrativa são características discriminatórias para classificar os *posts* com intenção racista [1].

A liberdade de expressão fornece a um indivíduo o compartilhamento de suas opiniões e crenças sobre qualquer coisa. No entanto, muitos usuários de mentalidade semelhante abusam da liberdade de expressão para fazer comentários ofensivos ou promover suas crenças que podem causar um impacto negativo na sociedade. A pesquisa mostra que esses indivíduos ou grupos de pessoas usam sites, redes sociais populares (Twitter e Tumblr) para tais atividades. Existem usuários que publicam comentários racistas visando grupos existentes, com base na análise do trabalho, foram definidos esses grupos em duas categorias: Religião e Raça.

A mineração de intenções de texto de mídia social é um problema tecnicamente desafiador devido à presença de *scripts* multilíngue, gramática incorreta, palavras erradas, texto curto, siglas, abreviaturas, sarcasmo e postagens baseadas em opinião. Foram encontrados exemplos de conteúdo ambíguo disfarçado que dificulta a classificação, mesmo para a anotação humana. O trabalho apresentado neste artigo é motivado pela necessidade de desenvolver um sistema para identificar automaticamente um *post* feito com intenção racista.

3.1.1 Configuração do experimento

Coleta de dados: o experimento foi conduzido em um conjunto de dados de código aberto e em tempo real extraído do microblog do Tumblr. Realizou-se uma inspeção manual onde foi criado um léxico das principais tags K que são comumente usadas por racistas ou grupos radicais. Por exemplo, (*islamophobia, islamemau, supremacy, blacklivesmatter*). Foi implementado um método de inicialização para criar o conjunto de dados para utilização desse léxico com tags de semente para a pesquisa no Tumblr. Para cada tag, extraiu-se somente postagens textuais (texto e citações) e estendido o léxico ao adquirir outras tags exclusivas das postagens extraídas. Foi executado o modelo até obter um número desejado de postagens ou até o modelo convergir (começar a extrair postagens duplicadas). No Tumblr, foi extraido um total de 3.228 mensagens de texto feitas por 2.224 blogueiros exclusivos consistindo de 10.217 tags únicas. Foi removido todos os posts duplicados e não ingleses dos dados onde foi disponibilizado publicamente para o experimento.

Anotações de dados: foi utilizado 2.456 postagens em inglês para anotação que cobre apenas 83% dos dados extraídos. Uma vez que, foi utilizado o método *bootstrapping* para coletar os dados, que extrai uma grande quantidade de mensagens barulhentas que não pertencem ao tópico definido (raça e religião). Portanto, primeiro foi anotado as postagens relacionadas ao tópico e depois rotuladas como intenções (racistas / radicais) ou desconhecidas. Foi empregado dois anotadores com 2 à 3 anos de experiência em usar o microblog do Tumblr afim de verificar os dados. Os resultados do tópico e a declaração de intenções realizadas em 2.456 postagens revela que 2.419 (292 tópicos e 2.127 desconhecidos) com o mesmo rótulo dos dois anotadores. Descartou-se as 37% postagens restantes como anotação inconsistente. Ambos os anotadores rotulam mais esses 292 tópicos como intenções ou desconhecidos. Os anotadores concordam em 278 postagens (103 intenções, 175 desconhecidas), enquanto há uma inconsistência nas 14 postagens restantes. O valor do coeficiente Kappa de Cohen para anotação tópica e intencional revelam que os anotadores concordam em mais de 90% do tempo.

Os resultados da anotação mostram que as postagens da intenção são apenas 37% das postagens do tópico e apenas 4% do conjunto de dados experimental completo, revelando que os dados rotulados são altamente desequilibrados.

3.1.2 Identificação das características

Modelagem de tópicos: a análise e anotação revela que, apesar de não ter certos termos-chave específicos do tópico, uma postagem pode ser uma mensagem intencional para o método de classificação baseado em palavras-chave, que não funciona com precisão e gera um grande número de falsos alarmes. Portanto, foi utilizado estatística e técnicas de processamento de linguagem natural para realizar a modelagem de tópicos em posts de Tumblr. Afim de diminuir os falsos alarmes foi utilizado o *Alchemy Taxonomy API* e *Tagging API* para identificar vários conceitos e classificar a postagem nas categorias de tópicos.

Sentimento e Análise de Tom: investigou-se a linguagem de uma narrativa, analisando vários tipos de sentimentos e traços de personalidade em uma publicação. Para isso foi realizado uma análise linguística de mensagens de texto usando *Alchemy Document Sentiment API* e *IBM Watson Tone Analyzer API*. A análise do sentimento identifica a polaridade positiva e negativa de uma postagem enquanto a análise de tons mede o nível de três cate-

gorias, incluindo tons emocionais, sociais e de escrita. Tom de emoções analisa o texto de uma postagem distribuindo em cinco emoções (alegria, medo, tristeza, raiva e desgosto. As tendências sociais analisam os traços de personalidade do texto (abertura, conscientização, extraversão, amabilidade e alcance emocional de uma narrativa). O tom de escrita identifica as pistas de linguagem do autor, medindo o estilo analítico, confiante e tentativo de escrita.

Etiquetagem semântica: a marcação semântica de uma publicação identifica o papel semântico de cada termo presente no conteúdo. Também identifica as frases ocultas que desempenham um papel importante na publicação. Nesse trabalho foi utilizado o Sistema de Análise Semântica UCREL para marcar cada publicação no conjunto de dados. Todas as tags semânticas em uma publicação são compostas por um rótulo de nível geral ou alto e um valor numérico que mostra a divisão de cada rótulo no léxico. Foi removido todas as pontuações e caracteres especiais do conteúdo etiquetados semânticamente e decodificado todos os termos restantes com seus respectivos rótulos em tags.

3.1.3 Classificação

A terceira fase é um classificador baseado em aprendizagem de conjunto em cascata, que consiste principalmente em duas etapas: classificação tópica e classificação intencional. Foi treinado o modelo a partir de vetores de recursos criados para classifição de posts com intenção racista.

Classificação do tópico: utiliza recursos linguísticos de modelagem tópica de cada publicação para identificar os tópicos. Foi tomada uma amostra aleatória de 50 postagens de 292 postagens anotadas como tópicos e extraiu-se sua taxonomia e conceitos do espaço de recursos. Dois léxicos independentes foram criados, desses conceitos, com tópicos rotulados que têm uma pontuação de confiança acima de 0.40. Houve uma filtragem manual da lista dessa taxonomia finalizando com seguintes rótulos: (religião e espiritualidade, sociedade / conflito e guerra, sociedade / racismo, sociedade / crime pessoal / crime de ódio, direito, governo e política / espionagem e inteligência / terrorismo e direito, governo e política / questões legais / direitos humanos).

Foi utilizado um método baseado em pesquisa para checar se a postagem pertence a qualquer uma dessas taxonomias e se tiver uma pontuação de confiança acima de 0.40. Se sim, então, ela é classificada isso como uma publicação tópica. Se uma postagem contém

uma ampla gama de taxonomias (> 5), identifica-se os principais conceitos K na postagem é classificada como tópico post com base na sua presença no léxico.

Classificação de intenção: a intenção de uma postagem não pode ser totalmente determinada apenas pela mineração das palavras-chave no conteúdo. Mas também requer entender e prever a tendência psicológica, os tons de sentimento e o idioma da narrativa. Isso também exige analisar o papel semântico das palavras-chave relacionadas ao tópico usadas na publicação. Foi realizado a classificação nas postagens de Tumblr, treinando o modelo em sentimentos, semântica e linguagem, com base em critérios de um texto. Em um nível alto, foi criado um espaço vetorial de 5 recursos definidos (F1: Sentimento do documento, F2: Marcação semântica, F3: Emoções, F4: Tom de escrita e F5: Tendências sociais), que é ainda categorizado em 15 vetores exclusivos. Foi definido a classificação intencional como um problema de classificação de uma classe. Portanto, os dados de treinamento contém apenas postagens de classe positiva (intenção). Foram implementados três diferentes classificadores de uma classe (Random Forest (RF), Naive Bayes (NB) e Decision Tree (DT)) e houve uma comparação da sua precisão para as postagens classificadas como tópicos no estágio 1.

O modelo foi treinado para cada classificador e realizou-se validação cruzada 5 vezes. Apenas 12% das postagens são rotuladas como mensagens de intenção, tornando o conjunto de dados experimental altamente desequilibrado.

3.1.4 Avaliação de desempenho

Com base nos resultados do acordo com o anotador, foi avaliado a Precisão do classificador comparando os resultados observados com a classe rotulada real. Realizou-se experiências em 2,419 postagens e o classificador de tópicos proposto classifica 346 postagens como classe alvo (tópico) e 2.073 posts como desconhecidos. Com uma classificação errada de 3.8% e 1.6% na identificação de postagens alvo e outliers (desconhecidas). Os resultados revelam que, para a classificação do tópico, foi alcançado uma precisão de 73% (253 / (253 + 93)) e um recall de 86% (253 / (253 + 39)).

Dado que dados são altamente desequilibrados e apenas 12% dos posts são rotulados como classe alvo (intenção), executou-se cada um dos classificadores (RF, NB e DT) usando uma validação cruzada 5 vezes. Como as medidas de precisão são tendenciosas na maioria,

foi avaliado o desempenho do classificador de intenção usando duas métricas de recuperação de informações padrão: Precisão e Área sob Curva do Receptor do Operador (AUC).

Devido à classificação errada na modelagem de tópicos, foi avaliado o desempenho da classificação da intenção em duas etapas. Foi executado o modelo em todas as 346 postagens classificadas como tópicos na etapa anterior e 253 posts Tumblr corretamente classificados como tópicos. Os resultados revelam que o classificador de intenções de uma classe fornece uma taxa de precisão mais alta para (Test-Data1) e a filtragem de posts não baseados em tópicos do conjunto de dados melhora ainda mais a precisão da classificação da intenção. Segundo os autores isso provavelmente está associado ao fato de que posts desconhecidos representam uma ampla gama de Sentimentos e sugestões de linguagem. O RF supera os algoritmos NB e DT e fornece a máxima precisão (0.78, 0.81) e recall (0.82, 0.84) para (Test-Data1) e (Test-Data2).

Os resultados mostram que as postagens incorretamente classificadas no estágio 1 provocam uma diminuição na classificação da precisão da intenção. Por fim os resultados ainda apresentam que se a taxonomia de uma postagem for desconhecida (Test-Data1), cada algoritmo tem uma probabilidade de aproximadamente 0.60 para classificá-la como mensagem intencional.

3.2 Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation

O aumento da presença online de *jihadistas* criou a possibilidade de indivíduos serem preconceituosos através da Internet. Até o momento, o estudo da radicalização violenta se concentrou em sites e fóruns dedicados a jihadistas. Os participantes nesses locais podem ser descritos como "mentes formadas". Rastrear uma plataforma global de redes sociais, como o YouTube, por outro lado, tem o potencial de descobrir conteúdos e interações que visam a radicalização daqueles usuários com pouco ou nenhum interesse prévio aparente no *jihadismo* violento.

Reuniu-se um grande conjunto de dados de um grupo dentro do YouTube que foi identi-

ficada como potencialmente radical. Analisou-se esses dados usando a análise de rede social e ferramentas de análise de sentimentos, examinando os temas discutidos e a polaridade do sentimento (positivo ou negativo) para esses tópicos. Em particular, o estudo concentrou-se nas diferenças de gênero neste grupo de usuários, sugerindo visões mais extremas e menos tolerantes entre usuários do sexo feminino.

Os *jihadistas* aumentaram significativamente sua presença online desde o 11 de setembro aumentando a possibilidade de exposição ao conteúdo de jihadista violento em linha, resultando em usuários individuais sendo radicalizados através da Internet.

A radicalização online é concebida como um processo pelo qual indivíduos, através de suas interações online e exposição a vários tipos de conteúdo da Internet, vêem a violência como um método legítimo de resolver conflitos sociais e políticos. Abordagens recentes para o estudo da radicalização violenta se concentraram em sites e fóruns de jihadistas dedicados. O presente trabalho baseia-se em pesquisas anteriores sobre os links entre o vídeo jihadista e a radicalização em linha, e a contribuição é uma análise detalhada de um conjunto de dados real do YouTube. Esta análise usa uma aplicação de análises de rede de sentimentos, lexicais e sociais, que permite examinar e caracterizar os usuários de fóruns radicalizados, com especial ênfase nas diferenças de gênero entre os usuários.

3.2.1 Configuração do Experimento

Ao analisar as informações do perfil do usuário do YouTube e a discussão iniciada pelos membros do grupo foi identificado, duas questões de pesquisa:

- Esse grupo foi povoado por radicais que estavam em posição de atrair outros para sua esfera de influência?
- Quais foram as diferenças, se houver, em termos de radicalidade entre o conteúdo publicado e as interações envolvidas pelos membros do grupo masculino e feminino?

A coleta e análise de dados consistiu em uma série de etapas: um rastreamento do YouTube para coletar dados relevantes, uma análise de rede desses dados e análise léxica do corpus para informar a análise do sentimento dos documentos reunidos.

3.2.2 Rastreamento do YouTube

Foi rastreado todos os comentários e perfis de usuários no grupo do YouTube, para coletar dados do mundo real de um fórum, os autores comentam que poderia se desenvolver uma agenda radicalizada. O grupo, o nome e outros detalhes foi omitido por motivos de privacidade, com mais de 700 membros. Um rastreamento automatizado dos dados do YouTube do grupo foi realizado pela primeira vez em 8 de janeiro de 2009 e novamente em 16 de janeiro de 2009.

O rastreamento inicial reuniu comentários dos usuários e o último coletou informações de perfil do usuário. Os comentários foram lidos da parte "Comentários do canal" dos perfis dos membros do grupo. As informações do perfil do usuário foram lidas para todos os membros do grupo e para qualquer usuário do YouTube que tenha postado no canal de um membro do grupo. O corpus utilizado neste documento é, portanto, composto de comentários de usuários e texto de perfil. No total, existem mais de 13.000 perfis e 122.011 comentários, gerando mais de 135.000 "documentos" únicos, cada um com um perfil de usuário ou um comentário de usuário. Os membros do grupo foram responsáveis por mais de 22.000 documentos, enquanto 13.000 outros usuários foram responsáveis por mais de 113.000 documentos. Dos 8.682 usuários que declararam seu gênero, 2.715 (31%) são do sexo feminino. Não foi identificada informações de gênero disponíveis para mais de 5.000 usuários. A idade proclamada dos membros do grupo variou de 14 a 107 anos.

3.2.3 Análise de sentimentos

A Análise de Sentimento foi realizada em todos os documentos coletados durante o rastreamento. Inicialmente, foi desenvolvido um sistema para a participação que foca na detecção de opinião e detecção de polaridade de opinião em um corpus de blog em larga escala. Diferenças na natureza dos dados, no entanto, tornaram os resultados não confiáveis.

No YouTube, a maioria dos comentários é consideravelmente inferior a 50 palavras de comprimento, muito mais curtas do que o comprimento do blog típico no corpus TREC de centenas de frases. Outra grande diferença foi que havia pouca evidência de subjetividade no texto do YouTube. Muitas vezes, quando um usuário do YouTube expressa uma opinião que eles simplesmente afirmam em vez de qualificá-lo com "Eu acho ..." ou "Eu sinto ...".

Esse comportamento não é visto no corpus do blog onde os autores estão interessados em distinguir a opinião do fato em suas postagens. Tudo isso significava que os componentes do sistema TREC projetados para detectar a subjetividade não foram úteis no contexto dos dados do YouTube e, portanto, dificultavam o desempenho.

Por estas razões, decidiu-se usar apenas o módulo de polaridade ("Módulo Lexicon") do sistema para obter documentos para positividade e negatividade. Este módulo usa um léxico, o SentiWordNet, que atribui valores de positividade e negatividade para sincronizar entradas no WordNet.

Como as orientações de polaridade dos termos não correspondem necessariamente à subjetividade, um documento pode simplesmente discutir um aspecto de um conceito usando termos polarizados, mas sem expressar opinião. Por exemplo, a frase “Ele está doente e cansado hoje”, tem dois termos orientados negativamente, “doente” e “cansado”, mas continua sendo uma declaração de fato negativa e não uma opinião negativa.

A análise de sentimento de execução em um documento gera uma pontuação positiva e uma pontuação negativa. Isso corresponde à orientação do termo médio no documento. Para o perfil de gêneros e conceitos, foi filtrado documentos por gênero de autor e termos de conceito, respectivamente. Para um conjunto de documentos, o índice de sentimento de positividade é definido como o escore de positividade médio para os documentos nesse conjunto, os escores de negatividade são calculado de forma semelhante.

3.2.4 Análise Lexical

Para extrair informações sobre os recursos lexicais do corpus, todos os comentários e perfis de usuários foram indexados pelo mecanismo de busca de Terrier. *Stopwords* foram removidos e o algoritmo de derivação de Porter foi usado para remover morfologias variantes de palavras. O índice do mecanismo de pesquisa foi utilizado para extrair várias estatísticas sobre o léxico do corpus, em particular informações sobre freqüências de uso de palavras sobre vários subconjuntos do corpus. As seguintes métricas são relatadas:

- Frequência de termo (TF): o número de vezes que o termo ocorre em toda a coleção.
- Frequência do documento (DF): número de comentários únicos ou perfis de usuários em que o termo ocorre.

- Frequência do usuário (UF): o número de usuários únicos que usam esse termo.

Os documentos nesta coleção são dominados por termos relacionados à religião. Os resultados dos 10 principais termos mais utilizados incluem 5 que se relacionam exclusivamente com a religião com os dois termos mais utilizados sendo **Allah** e o **Islã**. A discussão não é restrita ao islamismo, **Jesus** é o terceiro termo mais utilizado, embora pareça em relativamente em poucos documentos únicos com usuários únicos.

Comparando usuários masculinos e femininos, é impressionante que Jesus seja o termo mais utilizado pelos homens, embora não se registre no top 10 para mulheres. A menor frequência do usuário de Jesus para homens sugere, no entanto, que isso vem de um número relativamente pequeno de usuários que usam esse termo com muita frequência. Também é digno de nota que o único termo exclusivamente religioso no top 10 para mulheres é **Allah**.

3.2.5 Resultados da Análise Sentimental

A análise léxica do corpus permitiu determinar o que os usuários do grupo estavam falando, enquanto as técnicas de análise de sentimento foram capazes de nos informar sobre as opiniões ou atitudes dos usuários em relação a esses tópicos. Para determinar os tópicos alvo para os quais poderia extrair o sentimento do usuário, construiu-se "conceitos" de interesse potencial para os jihadistas dos 50 principais termos mais utilizados. Esses conceitos foram (América, Cristianismo, Islã, Israel, Judaísmo, Mubarak, Palestinae e Al-Qaeda). Para encontrar todos os documentos relevantes para esses conceitos, cada conceito foi expandido com variantes de ortografia, sinônimos, etc. Por exemplo, para o cristianismo, todos os documentos que continham qualquer um dos seguintes termos foram considerado relevante: **Jesus, cruz, cristão, bíblia**.

Observa-se que, para certos tópicos, como a América, o cristianismo e a Palestina, o nível de positividade e negatividade são amplamente similares. O tema mais exageradamente positivo para os homens é Mubarak, seguido do islamismo. Curiosamente, o judaísmo tem maior positividade do que a negatividade, enquanto o oposto pode ser dito pelo tema Israel, sugerindo que muitos desses machos são tolerantes com a religião judaica enquanto se opõem ao Estado Israelense.

Os resultados da análise do sentimento positivos para mulheres são Mubarak e o Islã. O

que é mais marcante, em termos da diferença entre os gêneros, é uma maior positividade feminina em relação ao tema Al-Qaeda e a maior negatividade feminina em relação ao tema do judaísmo. Isso sugere um ponto de vista mais extremo entre as usuárias do corpo: ao contrário dos homens, eles não parecem querer distinguir entre o Estado de Israel e a religião judaica. Eles também são muito mais positivos em relação à Al-Qaeda. O sentimento feminino negativo para o cristianismo também é forte, sugerindo novamente uma maior falta de tolerância de outras religiões. De fato, se a diferença entre a positividade e a nota de negatividade para cada conceito é considerada para cada gênero, a única inversão da polaridade é vista nos conceitos religiosos não-islâmicos, no cristianismo e no judaísmo. Para esses conceitos, os machos são mais positivos do que negativos e as mulheres são marcadamente mais negativas do que são positivos.

3.3 Classifying Racist Texts Using A Support Vector Machine

O trabalho apresenta uma visão geral das técnicas utilizadas para desenvolver e avaliar um sistema de categorização de texto para o projeto PRINCIP que se propõe a classificar automaticamente textos racistas. A máquina de vetor de suporte (SVM) é utilizada para classificar automaticamente páginas da web com base ou não racistas. PRINCIP é um projeto baseado em linguística que visa construir um sistema de classificação para páginas racistas na web através da análise baseada em corpus de conteúdo racista. Os padrões linguísticos identificados durante a análise de páginas da web podem ser formulados em regras e usados em um sistema de categorização para permitir a detecção de conteúdo ilícito na web.

A categorização de texto (TC) está relacionada com a atribuição automática de documentos a categorias predefinidas. O TC moderno toma emprestado e aplica muitas técnicas de dois campos de pesquisa estabelecidos: Recuperação de informação e Aprendizado de máquina.

Os métodos atuais de filtragem do racismo dependem fortemente de palavras-chave ou a rotulagem manual de material ofensivo. Para implementar sistemas de filtragem bem sucedidos, é necessário um esforço humano considerável, não apenas nos estágios iniciais de construção de filtros, mas também de forma contínua à medida que os alvos do racismo

mudam, à medida que o idioma evolui, os sites existentes são editados ou novos sites são adicionados. As técnicas automáticas de categorização de texto foram relatadas como bem sucedidas quando aplicadas a outros domínios, tais como categorização de notícias e tais métodos, levaram a grandes melhorias na produtividade, bem como economias em termos de tempo e mão-de-obra. Dada a fluidez do racismo na web, esta é uma área que pode se beneficiar na aplicação de técnicas automáticas de categorização de texto.

3.3.1 Detecção de textos racistas

Detectar o racismo na Internet não é apenas um problema baseado em tópicos, como na classificação de notícias, em vez disso, é mais parecido com a detecção de gênero, na medida em que não se esta realmente preocupado com o tema em si, mas tentando identificar características que discernem a atitude de um autor em relação ao tópico, algo que é ortogonal ao tópico real. Experimentos com PRINCIP revelaram que haveria diferenças em algumas distribuições lexicas, colocação e POS em documentos racistas e não racista. Com base nesta análise do domínio de textos racistas, decidiu-se comparar várias representações de recursos, verificando padrões de bagagem de palavras e bi-gramas de palavras na detecção de textos racistas. Um SVM foi treinado em cada representação para identificar o método mais produtivo e a representação para detectar o racismo.

3.3.2 Máquina de Vetor de Suporte

Foi utilizado o SVM para aprender os recursos dos conjuntos de treinamento e classificar novos documentos não vistos. Os SVMs são um método de aprendizagem muito poderoso que "desde a sua introdução já superou a maioria dos outros sistemas em uma ampla variedade de aplicações". Os SVMs superaram muitos dos problemas associados à eficiência do treinamento, como a superposição. Eles são capazes de se generalizar bem em grandes espaços dimensionais, no presente trabalho, onde há uma rica representação de palavras, bi-gramas, etc., o que significa que as soluções sempre podem ser encontradas de forma eficiente mesmo para conjuntos de treinamento com muitos exemplos. A representação compacta da hipótese que está sendo aprendida significa que a avaliação em entradas não vistas é muito rápida, tornando assim eficiente quando se trata de testes.

Dado um documento de entrada *d*, para chegar à classe de saída *c*, o SVM deve aprender a relação entre os emparelhamentos de entrada e saída. A função que faz isso é conhecida como a função alvo. Isso permite que a máquina tome uma decisão sobre o classe alvo dos documentos invisíveis.

3.3.3 Resultados

Durante o projeto PRINCIP, coletou-se um corpus de 3 milhões de palavras. O corpus foi dividido em conjuntos de dados de tamanhos variados com um número igual de documentos racistas e não racistas em cada conjunto.

Os resultados demonstram que o recall melhora à medida que o conjunto de treinamento aumenta. Um aumento constante foi relatado com os números de precisão / recall para o conjunto de dados final alcançando 92.55% / 87.00%, uma melhoria considerável na precisão / recall para o conjunto 1.

Para cada precisão de conjunto de dados, melhorou drasticamente em comparação com a representação da BOW enquanto os números de recall caíram entre 10-15%. À medida que o conjunto de treinamento é aumentado, a precisão diminui ligeiramente ao se recuperar melhora atingindo 75% no conjunto.

A precisão no conjunto de teste para a representação BOW superou bi-grams: para o conjunto 3 para a precisão no conjunto de teste para o BOW foi 87.33%, enquanto os bigrams obtiveram uma precisão de 84.77%.

Tanto a abordagem BOW como os bi-grams têm suas vantagens com BOW, resultando em recordações elevadas e bigrams dando alta precisão.

Embora isso possa ser computacionalmente caro, também seria interessante ver o que afeta a BOW e bigrams juntos como uma representação, teria precisão e recall.

É possível construir um sistema de classificação automática para a detecção de racismo na web. Os autores propõem como pesquisa futura a formação de SVMs para sequências de palavras de tri-gramas e parte das tags de fala - de modo a identificar o método mais efetivo que permitirá a classificação de documentos racistas na web.

3.4 A Lexicon-based Approach for Hate Speech Detection

A maior parte do trabalho sobre análise de sentimentos se concentra em domínios baseados em revisão, como revisões de filmes e produtos. Atualmente, o conteúdo gerado pelo consumidor (CGC) fóruns online, blogs e seções de comentários em sites de revisão de notícias são importantes fontes de opiniões das pessoas sobre uma variedade de questões atuais que vão desde finanças, educação, religião, política e uma série de questões sociais gerais. O discurso da web no domínio da análise de sentimento, inclui a avaliação de fóruns da web, grupos de notícias, e blogs. Este domínio, no entanto, tornou-se uma fonte comum de discussão, mensagens e novidades. Por exemplo, nos chamados fóruns da *deep web* (web profunda), extremistas e grupos terroristas se comunicam, compartilham ideologias e usam os fóruns como condutas para radicalização.

A tarefa de determinar sentimentos de discurso de ódio em fóruns, blogs e comentários em notícias, tem despertado vários pesquisadores para investigar a detecção de preconceito e mensagens violentas nas mídias sociais bem como a propagação de mensagens odiosas nos fóruns da web profunda. Diferente de outras formas de discurso de ódio que utilizam linguagem ofensiva e ameaçadora que visa certos grupos de pessoas com base em sua religião, etnia, nacionalidade, cor ou gênero. A fonte das mensagens de ódio geralmente é um membro de uma grupo supostamente rival, como pertencer a outra comunidade étnica. A disseminação de mensagens de ódio pode ser através de sites dedicados associados a um grupo coeso de membros, mas também pode ser através de sites populares, como Yahoo!, Twitter ou Facebook, onde questões atuais ou artigos de notícias podem provocar respostas com linguagem estereotipada.

O presente trabalho está preocupado com a tarefa de desenvolver um classificador que possa ser aplicado a detectação automatica do discurso de ódio nas mídias sociais. O principal desafio descrito é a falta de um corpus rotulado que possa ser aplicado diretamente a essa tarefa. Foi utilizado uma abordagem baseada na análise de subjetividade e desenvolvida classificações de discurso de ódio.

Com a utilização de técnicas de análise de sentimento de nível de sentença, foi iniciado a detecção de subjetividade onde foi utilizado uma abordagem baseada em regras para separar frases objetivas de frases subjetivas. Usando pistas de subjetividade aprendidas com o

Multi-Perspective Question Response (MPQA) corpus e outras fontes, foi treinado a base de regras do detector de sentença subjetiva. Como as pistas têm orientações para as opiniões, argumentações, e temas de polarização, eles são adequados para o problema sentimental.

Para construir o léxico de fala de ódio, iniciou-se a extração subjetiva apresentando caracteres de palavras semânticas que credenciam uma frase subjetiva. Uma vez que o domínio do discurso de ódio está fortemente carregado com dependência e contexto de domínio léxico específico, foi aumentado o léxico semântico subjetivo com corpus léxico gerado. Usando *bootstrapping* e WordNet, foi adicionado ao léxico desenvolvido, verbos relacionados ao ódio e padrões gramaticais gerados pelo tipo de dependência relativos a as três áreas temáticas identificadas. Com base no léxico, foi criado um discurso de ódio aplicado a detecção com três níveis de: "Não há ódio", "Fraco ódio" e "Forte ódio". Foi testado o aplicativo usando o corpus anotado consistindo de 500 parágrafos rotulados. O processo foi dividido nas seguintes quatro etapas:

- Passo 1: Abordagem de aprendizagem de regras para extrair frases subjetivas.
- Passo 2: Frases subjetivas identificadas no passo 1 acima, extraindo a semântica e características de palavras subjetivas.
- Passo 3: Usando *bootstrapping*, foi aumentado o léxico na etapa 2 com padrões nominais com base nas classes semânticas da religião, etnia e raça dos verbos relacionados ao ódio.
- Passo 4: Foi criado e testado o classificador com o corpus anotado com base no recursos identificados nas Etapas 2 e 3.

3.4.1 Corpus de dicurso de ódio

O corpus de discurso de ódio consiste em duas fontes diferentes que têm uma orientação profundamente diferente em termos de público-alvo e apresentação. Para a fonte principal, rastreou-se em datas diversas um total de 100 postagens de blog de 10 sites diferentes, 10 para cada site, a partir de uma lista fornecida no diretor de ódio. Este é um diretório compilado por Raymond Franklin de sites que são considerados geralmente ofensivos. Um desses

sites é *stormfront.org*, um site neonazista, considerado o primeiro grande "site de ódio" doméstico nos Estados Unidos.

Os blogs que foram tematicamente relacionados a áreas de etnia, religião e nacionalidade. A maioria desses sites fornece fóruns de discussão baseados em temas com inúmeras questões atuais, contribuição de ensaio dos membros, convidados e seção de comentários para os leitores comuns. A maioria das discussões são discursos intelectuais em áreas sutis, como ideologia e ciência. A linguagem de texto utilizada, apesar de ter algum viés subjacente, raramente é explicitamente prejudicial, mas contém insinuações contra grupos rivais percebidos.

O site de origem secundária consiste principalmente em um trecho de frases de parágrafo relacionadas ao conflito Israel/Palestina. O idioma usado aqui é mais direto e facilmente discernível mesmo por um leitor casual.

Dois estudantes de pós-graduação da universidade dos autores realizaram a anotação da amostra representando aproximadamente 30% do corpus. Para o primeiro corpus, aleatoriamente selecionaram 3 postagens de blog de cada um dos 10 sites para um total de 30 blogs. Da mesma forma para o segundo corpus, em um documento de 150 páginas, dividiu-se o documento em três seções de 50 páginas e foi selecionado as 15 primeiras páginas em cada seção para um total de 45 páginas. Para facilitar o processo de anotação, as avaliações foram feitas em uma base de parágrafo em caso de documentos semelhantes ao ensaio onde foi usado trechos de frases para as cotações. Foi avaliado o uso de uma escala de 3 pontos de Not-Hateful (NH, tanto positiva quanto neutra), débilmente odiosa (WH) e fortemente odiosa (SH). No total, 180 parágrafos foram rotulados para o primeiro corpus e 320 parágrafos para o segundo corpus.

3.4.2 Abordagem proposta

A tarefa é desenvolver um léxico de expressões de sentimento usando semântica e características de subjetividade com orientação para a fala de ódio e depois usar esses recursos para criar um classificador para detecção de fala de ódio.

A abordagem prossegue em três etapas principais. A primeira etapa envolve a subjetividade detecção e destina-se a isolar frases que tenham expressões subjetivas daqueles que geralmente expressam sentimentos objetivos. Como discurso de ódio, emoções, opiniões,

avaliações e especulações estão carregadas de expressões fortemente subjetivas. O benefício direto de usar frases subjetivas é facilitar a detecção dessas opiniões. Na segunda etapa, construiu-se um léxico de palavras relacionadas ao ódio usando um método baseado em regras usando recursos subjetivos identificados a partir das frases e características semânticas aprendidas diretamente do corpus. Então, na fase final, um classificador que utiliza recursos criados a partir do léxico desenvolvido foi utilizado para testar o discurso de ódio em um documento.

3.4.3 Análise de subjetividade

Para construir um classificador de subjetividade tanto as abordagens baseadas em regras como as abordadas pela aprendizagem foi utilizado os métodos baseados em regras que não envolvem aprendizado e geralmente dependem de um lista pré-compilada ou um dicionário de pistas de subjetividade. A utilização de abordagens baseadas na aprendizagem, e algoritmos de linguagens de máquina em corpos marcados e sem etiqueta foram utilizados para aprender padrões ou outras pistas subjetivas.

3.4.4 Agregando Opiniões para Detecção de Discurso de ódio

Com base no léxico criado, os recursos que são de interesse para a aplicação são típicas palavras de opinião negativa descritas acima como polaridade negativa e palavras odiosas que são relacionados ao ódio, mas não fazem parte do léxico de polaridade negativa. Além disso, há características inspiradas em padrões gramaticais que foram capturadas como características de duas palavras de co-ocorrência e foi representada contexto e recursos dependentes do domínio. Para fazer uma previsão de uma sentença se é odiosa ou não, os autores utilizaram uma avaliação de nível de sentença orientada pelo número e confiabilidade de palavras de opinião em uma frase, bem como ocorrências de outras expressões lexicais em sentença.

Os principais desafios foram combinar diferentes recursos, bem como capturar o contexto para os recursos com base em temas de co-ocorrência. O algoritmo permite variar o impacto dos vários conjuntos de recursos ao ter o ódio, o léxico e os recursos baseados em tema dependem dos recursos semânticos. Cada sentença é classificada como subjetiva do coro de anuário anotado é lido e a base da palavra em cada frase atribuída uma etiqueta POS.

Cada palavra marcada com POS é pesquisada em várias categorias de léxicos. Uma frase é classificada em categorias de ódio com base no seguinte conjunto de regras:

- Todas as palavras de opinião negativas em uma frase são identificadas primeiro. Se dois ou mais as palavras marcadas como fortemente negativas aparecem em uma frase, prevendo que a sentença seja fortemente odiosa. Caso contrário, se apenas uma única palavra aparecer como fortemente negativa é previsto a sentença como débilmente odiosa.

- Se apenas uma única palavra em uma frase for marcada como fortemente negativa, mas uma ou mais as palavras aparecem no léxico do ódio, então foi previsto que a sentença seja fortemente odiosa. Mas se apenas uma palavra aparece do léxico do ódio sem outro da categoria semântica, prevê-se que seja ódio.

- Se o par dependente do governador dos padrões de co-ocorrência incluirem um tema baseado em substantivo e uma palavra marcada como fortemente negativa ou uma palavra que aparece no ódio léxico, a sentença é fortemente odiosa. Se o substantivo temático aparecer ao lado de um palavra designada como fracamente negativa, prevê-se como odiosa.

3.4.5 Configuração do Experimento

No presente trabalho um sistema baseado no léxico proposto e algoritmos descritos acima foi implementado usando a linguagem de programação Java. A avaliação da eficácia do sistema na previsão de orações odiosas e classificação foram feitas. Para realizar a avaliação, o sistema consiste em módulos referentes a diferentes categorias de conjuntos de recursos, conforme determinado pelo léxico. Um usuário pode decidir usar características de polaridade negativa sozinhas, polaridade negativa combinada com características de ódio ou alternativamente, uma combinação envolvendo os dois conjuntos de recursos e, adicionalmente, o tema recursos baseados. Foi apresentado que o uso das três categorias assegurando a melhor previsão por frases detestáveis.

A entrada para o sistema é um corpus de texto e um léxico de pistas de subjetividade. Depois do pré-processamento básico, opcionalmente, o corpus textual passa pelo módulo de subjetividade e análise e é transformado em uma série de frases subjetivas.

Com base no algoritmo incremental, foi realizado experimentos da seguinte maneira:

- No primeiro experimento, utilizou-se as pistas de polaridade negativa extraídas. Para este experimento, foi usado toda a lista de verbos, substantivos, advérbios e adjetivos que são marcados com uma confiabilidade fortemente subjetiva e uma polaridade prévia negativa. Para gerar o léxico negativo de polaridade uma sentença foi julgada odiosa se conter pelo menos duas palavras com etiquetas negativas, uma das quais deve ser fortemente negativo.

- No segundo experimento, foi incluído os verbos "ódio"que não estão na lista fortemente subjetiva. Foi avaliado sua importância como léxico equivalente de polaridade negativa. No total, foram adicionados 73 novos verbos, incluindo palavras como matar e expulsar.

O experimento final, incluiu o léxico das características derivadas da padrões gramaticais baseados em temas. Por este experimento também foi incluido todas as 1289 palavras no SUJEITO que são marcadas com uma confiabilidade subjetiva e uma polaridade prévia negativa. Foi extrado dois padrões de ocorrência de co-ocorrência que satisfazem as condições no tema de recursos baseados. Somente padrões que ocorrem com uma freqüência de pelo menos duas vezes no corpus chegou à lista. No total, extraiu-se 103 desses padrões.

3.4.6 Avaliação e Resultados

O esquema de avaliação, como a implementação, é incremental e se concentra nos efeitos dos vários conjuntos de recursos. A avaliação foi baseada no número total de frases classificadas em "ódio forte", "ódio fraco"e "Não odeio"no corpus anotado descrito anteriormente. Os resultados apresentam o desempenho do sistema na previsão de frases "ódio forte". Os autores discutem que utilizando características semânticas com base apenas na polaridade negativa o desempenho é inferior a 70% tanto em recall e precisão tanto para o primeiro como no segundo corpus. No entanto, quando foi incluído verbos de ódio, os resultados de precisão aumentaram ligeiramente acima de 70% para ambos os corpus. Foi observado que o aumento foi mais notável com o primeiro do que no segundo corpus, isso pode ser explicado pela maior explicitação do primeiro corpus do que o segundo. O uso do comando e a linguagem orientada a ação está mais associada aos verbos que outras formas de fala. No

entanto, o impacto geral do léxico de verbo de ódio foi limitado, em grande parte porque a maioria dos verbos gerados já fazem parte do léxico de polaridade negativa.

Para verificar a eficácia do uso de frases subjetivas, foi comparado os resultados usando todas as frases (objetivas e subjetivas) do corpus anotado. Como frases subjetivas, o processamento extrai características semelhantes do léxico. Os resultados apresentados demonstram uma queda substancial na precisão e no recall para ambos os corpus.

Pesquisas anteriores sobre detecção de fala de ódio concentraram-se principalmente na determinação de racistas textos em um documento. Neste trabalho, foi discutido e estendido o domínio do discurso de ódio para inclusão etnia e religião. Embora a comparação dos aspectos étnicos e religiosos seja impossível devido à falta de conjunto de dados unificadores, foi utilizado o conjunto de dados para comparar o desempenho do método com uma abordagem supervisionada empregada para prever tweets racistas. Sua abordagem emprega uma abordagem de classificação Naïve Bayes para a previsão de tweets como racistas ou não racistas.

3.5 Comparação entre trabalhos e proposta

Com os trabalhos descritos nas seções anteriores, foi desenvolvida uma análise comparativa entre os trabalhos e a proposta desenvolvida. Foi destacado alguns aspectos em alguns trabalhos, que podem ser visualizados na tabela 3.1 abaixo.

O trabalho de [29] é discutido a detecção de linguagem abusiva no conteúdo online gerado pelo usuário. Neste trabalho, foi desenvolvido um método baseado em aprendizado de máquina para detectar discurso de ódio em comentários de usuários online. Foi desenvolvido um corpus de comentários de usuários anotados para linguagem abusiva, o primeiro de seu tipo. Por fim foi utilizado, a ferramenta de detecção para analisar o idioma abusivo ao longo do tempo e em diferentes configurações para aprimorar ainda mais o conhecimento desse comportamento

Já em [1], [10], [18] e[17] apresentam propostas semelhantes na identificação de comentários dos usuários que usam sites populares para postar discurso preconceituoso contra várias religiões e raças. Para solucionar esse problema os autores apresentam propostas de desenvolvimento de um classificador de aprendizado de máquina para identificar os posts

com intenção racista.

A abordagem desenvolvida neste trabalho utiliza tanto os algoritmos de aprendizagem de máquina como também foi desenvolvido um dicionário léxico com termos preconceituosos, os demais trabalhos apresentados utilizam em suas abordagens unicamente com dicionários juntamente com a análise de sentimentos ou a aprendizagem de máquina. No nosso trabalho combinamos ambas as abordagens afim de verificar sua aderência com o presente tema.

Tabela 3.1: Comparação entre abordagens e a proposta

PROPOSTAS	**Agarwal et al., (2016)**	**Bermingham et al., (2009)**	**Greevy et al., (2004)**	**Gitari et al., (2015)**	**Nobata et al., (2016)**	**Silva Neto (proposta)**
Abordagem baseado em Aprendizagem de Máquina	Não	Não	Sim	Não	Sim	Sim
Abordagem baseada em dicionário léxicos	Sim	Sim	Não	Sim	Não	Sim
Pré-processamento	Sim	Sim	Não	Não	Sim	Sim
Análise de sentimento	Sim	Sim	Não	Sim	Sim	Sim
Classificação	Sim	Sim	Sim	Sim	Sim	Sim
Extração de conteúdo	Sim	Sim	Sim	Sim	Não	Sim

O próximo capítulo abordara a metodologia do trabalho, discriminando todos passos que foram executados para o desenvolvimento da abordagem.

Capítulo 4

Metodologia

O presente trabalho foi dividido em quatro etapas: coleta da base de dados, pré-processamento, identificação de termos preconceituosos e classificação. Na figura 4.1 é apresentado as etapas do desenvolvimento. Para o nosso trabalho, empregamos um método de classificação supervisionado que utiliza um conjunto de características de PLN que medem diferentes aspectos do comentário do usuário.

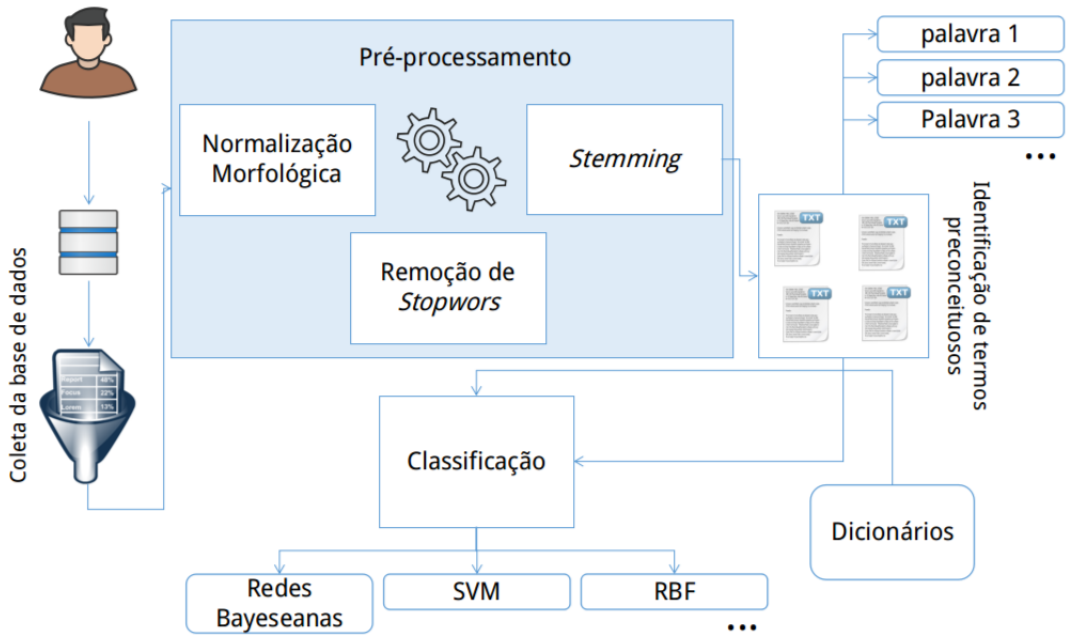

Figura 4.1: Passo a passo do desenvolvimento da proposta.

4.1 Coleta da Base de Dados

Esta etapa diz respeito a criação da base de dados utilizadas neste trabalho. A base de dados foi retirada através de uma filtragem manual do Twitter, o procedimento para essa filtragem, se deu a partir da definição de termos como: "preconceito", "racismo", "homofobia". A base contêm 500 tweets, ela foi dividida em, 200 tweets com sentenças preconceituosas e 300 tweets com sentenças neutras, é interessante destacar que a base está desbalanceada, devido a quantidade de tweets com conteúdo implícitos abusivos.

Essa rotulação foi criada por dois analisadores, quando houve divergência entre eles um terceiro analisador foi consultado, no período entre 02/07/2017 à 02/08/2017.

Dentre os tweets selecionados preoconceituosos foi levado em consideração comentários preconceituosos implíticos e explícitos apresentamos, por exemplo:

- "Quatro pretos tentaram pegar um táxi, mais de dez passaram, nenhum parou?"
- "Lugar de negro é na África. Não vem poluir o Brasil não".

Já entre os tweets neutros:

- "Diga não ao racismo, somos todos humanos."
- "Sabemos que não é uma tarefa fácil tirar a ideia de preconceito de alguém,mas vamos lutar pelo nosso país."

Houve dificuldade em encontrar esses tweets, pois muitos deles normalmente são excluídos quando publicados pela próprio rede ou denunciada pelos usuários.

4.2 Pré-Processamento

Nesta etapa a ferramenta divide as sentenças para serem avaliadas. Cada sentença passa por um pré-processamento onde serão eliminadas palavras com acentos, cedilhas e com pouco valor para o texto. As etapas do pré-processamento realizado são apresentadas na figura 4.2.

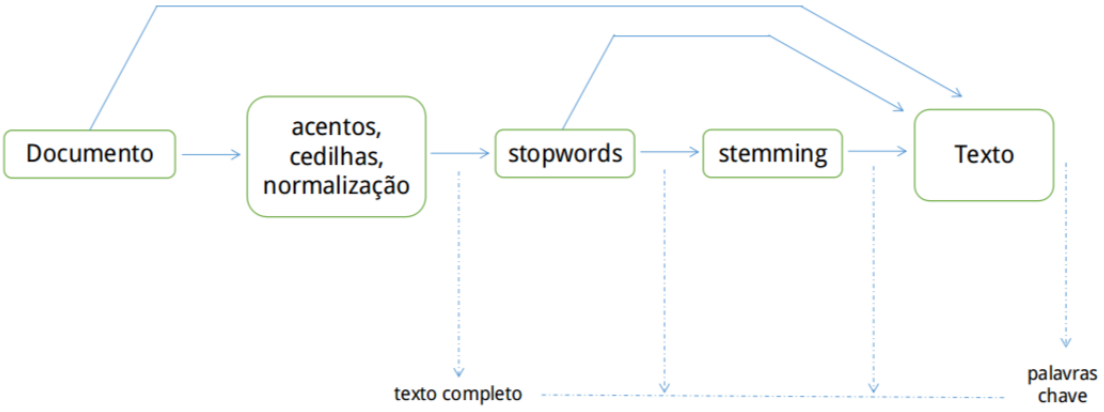

Figura 4.2: Visão lógica de um documento através das fases de pré-processamento de texto Baeza-Yates, R., Ribeiro-Neto, B. (2013).

1. **Normalização Morfológica**

   O objetivo principal da normalização do texto é facilitar a identificação dos termos, para isso, são eliminados palavras com acentos, com cedilhas, como também todo o texto é convertido passando as palavras maiúsculas para minúsculas.

2. **Remoção de *StopWords***

   Permite a eliminação de algumas palavras que não devem ser consideradas no documento, conhecidas como *stopwords*. Elas são consideradas não relevantes na análise de textos e incluem, normalmente, preposições, pronomes, artigos, advérbios, e outras classes de palavras [26].

3. ***Stemming***

   É uma técnica de identificação de radicais, permite a redução de uma palavra ao seu radical (ou raiz). Além da eliminação dos prefixos e sufixos, características de gênero, número e grau das palavras são eliminados. Isso significa que várias palavras acabam sendo reduzidas para um único termo [49]. A Figura 4.3, representa o passo a passo desse processo.

   Um exemplo para representação desse processo pode ser observado a partir da seguinte frase:

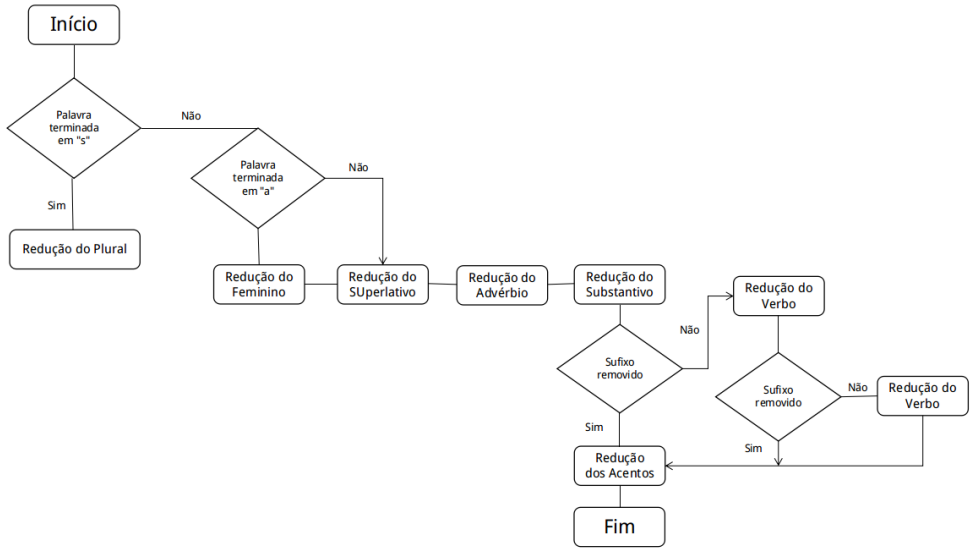

Figura 4.3: *Stemming* na Língua Portuguesa Morais, E. A. M., Ambrósio, A. P. L. (2007).

1) Primeiro passo, sem nenhum tipo de pré-processamento: "O único preto que eu gosto é você."

2) Segundo passo, retirada das *stopwords*: "único preto gosto você."

3) Terceiro passo, normalização morfológica: "unico pret gost voce."

4.3 Identificação de Termos Preconceituosos

Uma das principais contribuições do trabalho proposto foi a identificação dos termos preconceituosos para o idoma português brasileiro. Essa etapa foi dividida em dois passos: (i) criação de um dicionário léxico para o português; (ii) análise das palavras mais utilizadas no banco de dados proposto.

Para a primeira etapa o trabalho se baseou nos dicionários propostos em [34] e [43]. Como já foi dito, o primeiro é um dicionário geral para análise de sentimento em português. Dele foram retirados os grupos de palavras que podem indicar comentários preconceituosos.

Esse dicionário pode ser definido como uma base lexical para Língua Portuguesa, ele contêm quinze grupos de palavras divididas em: adjetivos e substantivos, ambas com para reconhecimento de expressões de emoções. Para o trabalho aqui desenvolvido foram selecionados os grupos com emoções positivas e negativas que mas se relacionavam com precon-

ceito, os demais grupos representavam emoções neutras. O dicionário contém no total 289 palavras, mas para o desenvolvimento deste trabalho foram selecionadas trinta palavras com emoções positivas e trinta palavras com emoções negativas respectivamente.

O segundo trabalho apresenta um recurso léxico para identificação de comentários racistas em alemão. Ele foca principalmente em palavras relacionados a raça, nacionalidade e religião. Neste caso, foi realizado uma tradução e adequação dos termos em alemão para o português. É importante frisar que nesse dicionário as palavras são categorizadas como termos racistas e termos neutros divididos em dezesseis grupos. Para o trabalho foram selecionados 412 palavras neutras e 683 palavras racistas. Na tabela 4.1 e 4.2 pode ser visualizado em maiores detalhes essas informações.

Tabela 4.1: Etapa 1: dicionário Pasqualotti

Dicionário 1 (Pasqualotti)		**Etapa 1 (selecionados)**	
Quantidade de palavras	289	Quantidade de palavras	60
Quantidade de grupos	15	Quantidade de grupos	6

Tabela 4.2: Etapa 2: dicionário Tulkens

Dicionário 2 (Tulkens)		**Etapa 2 (selecionados)**	
Quantidade de palavras	6000	Quantidade de palavras	1095
Quantidade de grupos	16	Quantidade de grupos	6

Na tabela 4.3 é apresentado alguns exemplos de palavras do dicionário de Tulkens [43], para acessar todas as palavras geradas a partir desses dicionários acessar o link[1].

[1]https://drive.google.com/open?id=0B1nWysJxwUHIUkpUdWZKdjNTU00

Tabela 4.3: Dicionário com termos neutros e racistas

iemenita	mexicano	minorias étnicas	nacionalista
iemenitas	moldova	imigrantes	não ocidental
jesuíta	muçulmano	americano	nacional
Camarões	extremista islâmico	requerente de asilo	remessas criminais
cantonês	terrorista islâmico	nativo	supranacionais
Cazaquistão	moçambicano	belga	terrorista
Kenya	Namíbia	nacionalidade estrangeira	turco
Quênia	nepal	estrangeira nativa	autorização de residência
quenianos	nepalês	comunista	refugiado
Quirguistão	Nicarágua	política de defesa	estranho
curdo	nova Zelândia	diplomático	desconhecido
kuwait	novo grego	expatriado	marrom
kuwaitianos	nigeriano	automobilismo	preto
congolês	norueguês	europeu	vermelho
cabeçalho	Uganda	expat	branco
coreano	ucraniano	trabalhadores convidados federais	Afeganistão
croata	ucrânia	ilegal	África
conterrâneo	coligação política oriental	é ilegal	africano
latina	timorenses	imigrante	Alaska

Após a criação dos dicionários foi realizado um estudo sobre as palavras mais importantes que aparecem na base de dados. Para isso, foi realizado todos os pré-processamentos propostos na seção anterior e as palavras foram pontuadas de acordo com sua frequência.

Para [26] nem todas as palavras presentes em um documento possuem a mesma importância. Os termos mais frequentemente utilizados (com exceção das *stopwords*) costumam ter significado mais importante, assim como as palavras constantes em títulos ou em outras estruturas, uma vez que provavelmente foram colocadas lá por serem consideradas relevantes ou descritivas para a ideia do documento.

Sendo assim, o cálculo de relevância de uma palavra em relação ao texto pode basear-se na importância da mesma. As análises baseadas em frequência costumam ser as mais utilizadas por serem mais simples, o grau de relacionamento de uma palavra com um texto dá-se o nome de peso.

Existem várias fórmulas para cálculo de peso, porém neste trabalho será utilizadod: TF-IDF, por melhor se adaptar aos objetivos da proposta.

- Frequência absoluta (TF)

Representa a medida da quantidade de vezes que um termo aparece em um documento. Onde, F é a frequência de um termo, *w* é um termo do texto analisado que é divido por *TodasPalavras* do texto.

$$TF(w) = \frac{F(w)}{F(TodasPalavras)}$$

- Frequência inversa de documentos (IDF)

A fórmula de frequência inversa de documentos (*inverse document frequency - IDF*), é capaz de aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos, justamente pelo fato dos termos de baixa frequência de serem, em geral, mais discriminantes [26].

$$IDF(w) = log\left(\frac{(N)}{n}\right)$$

O resultado TF-IDF é calculado da seguinte forma:

$$TF - IDF(w) = TF(w) * IDF(W)$$

Além desses cálculos de peso, serão utilizadas medidas para verificar, quantidade de palavras erradas no texto, tamanho de sentenças, quantidade de parágrafos e quantidade de erros de concordância.

Para a classe de texto preconceituoso as palavras que se repetiram várias vezes foram: **racista, negro, preto, cabelo, gay, peso e branco**. Por outro lado, as palavras de postagens neutras mais frequentes são: **racismo, preconceito, Brasil, luta, homofobia, cor e mulher**. Essas palavras também foram utilizadas como características.

4.4 Classificação

A primeira etapa da classificação foi a definição das características dos textos que seriam usadas. Para realizar uma avaliação mais abrangente foram definidos diferentes conjuntos de características:

- **Todas as Palavras (TP)**: todas as palavras do texto;
- **Palavras Frequentes da Classe Preconceituoso (PFCP)**: palavras que foram mais vezes utilizadas nas postagens consideradas preconceituoso;
- **Palavras Frequentes da Classe Neutra (PFCN)**: palavras que foram mais vezes utilizadas nas postagens consideradas neutras;
- **Dicionário 1 (Dic1)**: palavras consideradas preconceituosas do dicionário de [34];
- **Dicionário 2 (Dic2)**: palavras consideradas preconceituosas do dicionário de [43];
- **Dicionário 3 (Dic3)**: palavras consideradas neutras do dicionário de [43].

O primeiro conjunto de características utilizado foi o TP. Para isso, não foi realizado nenhum processamento no texto. Depois foram criados diferentes conjuntos de características utilizando o método MFD [35] nos conjuntos PFCP e PFCN. Esse método propõe a utilização de X palavras mais relevantes de cada postagem como característica, onde o valor de X é um inteiro, os autores mostraram que o valor ideal de X sempre está entre 1 até 10 para textos grandes. Como estamos trabalhando com textos extraídos do twitter ou sentenças, vão ser testados os valores de 1 a 3.

Por exemplo, quando o valor de X for definido como 1, vai ser recuperado o termo mais importante de cada texto para compor o vetor de características. Dessa forma, teremos os seguintes conjuntos de características:

- **TP**: com todas as palavras;
- **MFD1**: com uma palavra mais importante de cada texto e as palavras dos dicionários;
- **MFD2**: com duas palavras mais importantes de cada texto e as palavras dos dicionários;
- **MFD3**: com três palavras mais importantes de cada texto e as palavras dos dicionários;

Após a definição das características foram utilizadas técnicas de aprendizagem de máquina para identificar os textos com indícios de preconceito. Mais especificamente foram utilizados cinco classificadores de diferentes tipos: Redes Bayseanas, KNN, SVM, Árvore de Decisão e Rede Neural (RBF). Todos os classificadores foram utilizados usando a ferramenta WEKA (Waikato Enviroment for Knowledge Analysis) é uma suite de algoritmos de aprendizagem de máquina para tarefas de mineração de dados.

Os algoritmos podem ser aplicados diretamente a um conjunto de dados. Weka contém ferramentas para pré-processamento de dados. Weka contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. [48].

Capítulo 5

Estudo de Caso

No Brasil o Exame Nacional do Ensino Médio (ENEM) é uma avaliação do sistema de ensino implantada em nível nacional no país, sendo uma ferramenta de auxílio ao Ministério da Educação (MEC) para elaborar políticas de melhoria do ensino escolar. O ENEM permitiu a unificação das provas de vestibular e o processo de seleção e ingresso nas universidades federais brasileiras a partir do Sistema de Seleção Unificada (SISU). É importante destacar com a ampliação na utilização da nota ENEM como instrumento para o ingresso no ensino superior, o número de estudantes que passou a fazer essa prova aumentou significativamente- que, a cada ano o quantitativo de inscritos tem ultrapassado a escala de milhões de candidatos [20].

Uma das etapas do exame é a produção de uma redação, do tipo dissertativa-argumentativa, sobre um tema de ordem social, científica, cultural ou política. Para tanto, o MEC desenvolveu métricas para avaliação dessas redações que estão dispostas no Guia do Participante do ENEM [19], nele são apresentados cinco competências:

- **Competência 1** - demonstrar domínio da modalidade escrita formal da Língua Portuguesa, dissertativo-argumentativo em prosa;
- **Competência 2** - compreender a proposta da redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto;
- **Competência 3** - selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;

- **Competência 4** - demonstrar conhecimento dos mecanismos linguísticos necessários para construção da argumentação;

- **Competência 5** - elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Dentro das competências citadas na correção de redações do ENEM a **competência 5** busca analisar, entre outros aspectos, se o autor está respeitando os direitos humanos. Em outras palavras, o avaliador deve identificar se algum comentário preconceituoso foi escrito na redação. Devido ao grande aumento de participantes o custo para correção das redações também aumentou. Diante disso, é necessário propor soluções automáticas para auxiliar avaliadores na correção das redações.

5.1 Estudo quantitativo

Diante do exposto, foi conduzido dois estudos: quantitativo e qualitativo, para avaliar como a abordagem se comportaria neste cenário, sendo assim, foram selecionados classificadores que foram avaliados utilizando a base de dados descrita na seção anterior e as métricas tradicionais da literatura [7]:

- **Precisão**: avalia a quantidade de instâncias que foram classificadas corretamente;

- **Cobertura**: avalia a porcentagem instâncias de uma determinada classe que não foi classificada como pertencente a essa classe;

- ***F-Measure***: É uma média harmônica entre Precisão e Cobertura, como mostrado na Fórmula 5.1.

$$\textit{F-Measure} = 2 \times \frac{Preciso \times Cobertura}{Preciso + Cobertura} \tag{5.1}$$

Além disso, no processo de avaliação foi aplicado o método de validação cruzada com 10 k-fold [4] que divide o conjunto de dados em treinamento e testes, possibilitando uma melhor avaliação do classificador.

A Tabela 5.1 apresenta os resultados detalhados por algoritmo de classificação e grupo de característica utilizados.

Tabela 5.1: Resultados dos Algoritmos

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	TP	78,80	78,60	78,60
RBF	TP	59,30	46,40	49,50
SVM	TP	79,10	79,00	79,00
KNN	TP	69,70	59,50	58,20
Árvore de Decisão	TP	74,30	74,10	74,20
Redes Bayseanas	MFD1	77,20	77,40	77,30
RBF	MFD1	62,20	61,30	61,70
SVM	MFD1	80,20	80,40	80,20
KNN	MFD1	78,50	78,60	78,50
Árvore de Decisão	MFD1	79,60	79,20	79,30
Redes Bayseanas	MFD2	79,70	79,80	79,40
RBF	MFD2	65,60	61,10	63,30
SVM	MFD2	82,10	82,20	82,10
KNN	MFD2	79,20	78,20	78,40
Árvore de Decisão	MFD2	79,50	79,40	79,40
Redes Bayseanas	MFD3	80,00	80,00	80,00
RBF	MFD3	63,20	61,10	62,30
SVM	MFD3	80,80	80,80	80,80
KNN	MFD3	80,40	78,00	79,20
Árvore de Decisão	MFD3	79,30	79,20	79,20
Redes Bayseanas	Token N-Grams	83,60	83,80	83,60
RBF	Token N-Grams	41,60	64,50	50,05
SVM	Token N-Grams	80,08	81,00	80,09
KNN	Token N-Grams	79,40	79,00	79,20
Árvore de Decisão	Token N-Grams	75,70	75,60	75,70
Redes Bayseanas	Character N-grams	83,06	83,08	83,06
RBF	Character N-grams	45,90	62,50	50,02
SVM	Character N-grams	80,07	80,08	80,07
KNN	Character N-grams	78,40	77,00	77,20
Árvore de Decisão	Character N-grams	75,70	75,60	75,70

No primeiro cenário os resultados da combinação dos algoritmos com a característica TP que apresentaram os melhores resultados foram SVM (SMO) com 79,10% e Redes Bayseanas com 78,60% respectivamente. Já o pior resultado foi a combinação do RBF com TP apresentando 49,50%. Foi definido neste cenário todas as palavras do texto (TP).

Tabela 5.2: Resultados cenário 1

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	TP	78,80	78,60	78,60
RBF	TP	59,30	46,40	49,50
SVM	TP	79,10	79,00	79,00
KNN	TP	69,70	59,50	58,20
Árvore de Decisão	TP	74,30	74,10	74,20

No segundo cenário como resultados positivos na combinação do algoritmo com a caracterísitca MFD1, destacam-se SVM (SMO) e Árvore de decisão (J48) com 80,20% e 79,30% respectivamente. O pior resultado foi RBF mais MFD1 com 61,70%. Foi considerado nesse cenário a tupla MFD1 (uma palavra mais frequente de cada texto (PFCP e PFCN) + uma palavra dos dicionários (Dic1, 2 e 3).

Tabela 5.3: Resultados cenário 2

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	MFD1	77,20	77,40	77,30
RBF	MFD1	62,20	61,30	61,70
SVM	MFD1	80,20	80,40	80,20
KNN	MFD1	78,50	78,60	78,50
Árvore de Decisão	MFD1	79,60	79,20	79,30

No terceiro cenário destaca-se como positivo na combinação algoritmo mais caracterísitca MFD2, SVM (SMO) e Redes Bayseanas com 82,10% e 79,40% respectivamente. Com o pior resultado RBF com 63,30%. Foi considerado nesse cenário a tupla MFD2 (duas palavras mais frequentes de cada texto (PFCP e PFCN) + duas palavras dos dicionários (Dic1, 2 e 3).

Tabela 5.4: Resultados cenário 3

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	MFD2	79,70	79,80	79,40
RBF	MFD2	65,60	61,10	63,30
SVM	MFD2	82,10	82,20	82,10
KNN	MFD2	79,20	78,20	78,40
Árvore de Decisão	MFD2	79,50	79,40	79,40

No quarto cenário e último a combinação algoritmo com caracterísitca MFD3, novamente destaca-se positivamente o SVM (SMO) e Redes Bayseanas com 80,80% e 80,00%. Como pior resultado novamente o RBF com 62,30%. Foi considerado nesse cenário a tupla MFD3 (três palavras mais frequentes de cada texto (PFCP e PFCN) + três palavras dos dicionários (Dic1, 2 e 3).

Tabela 5.5: Resultados cenário 4

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	MFD3	80,00	80,00	80,00
RBF	MFD3	63,20	61,10	62,30
SVM	MFD3	80,80	80,90	80,80
KNN	MFD3	80,40	78,00	79,20
Árvore de Decisão	MFD3	79,30	79,20	79,20

No quinto e no sexto cenário foi implementado a feature Token-Ngrams e Character N-Gramas respectivamente, considerando a frequência de caracteres em palavras em português, sendo definido a média em (4), no Token-NGrams considerou apenas unigramas, no Character-NGrams foi considerado bigramas. Destacamos o resultado do Redes Bayseanas com 83,60%, em ambos os cenários esse classificador apresentou melhores resultados que o SVM (SMO).

Tabela 5.6: Resultados cenário 5

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	Token N-Grams	83,60	83,00	83,60
RBF	Token N-Grams	41,60	61,10	62,30
SVM	Token N-Grams	80,08	80,90	80,80
KNN	Token N-Grams	79,40	79,00	79,20
Árvore de Decisão	Token N-Grams	75,70	75,60	75,70

Tabela 5.7: Resultados cenário 6

Algoritmo	**Características**	**Precisão**	**Cobertura**	**F-Measure**
Redes Bayseanas	Character N-grams	83,06	83,08	83,06
RBF	Character N-grams	45,90	62,50	50,20
SVM	Character N-grams	80,07	80,08	80,07
KNN	Character N-grams	78,40	77,00	77,20
Árvore de Decisão	Character N-grams	75,70	75,60	75,70

5.2 Estudo qualitativo

Para avaliar a abordagem proposta no contexto educacional foi realizado uma análise qualitativa utilizando redações extraídas do Banco de Redações da UOL que são corrigidas seguindo os parâmetros estabelecidos pelo ENEM. Foram selecionadas cinco (5) redações para o experimento. Como as redações com comentários preconceituosos são eliminadas da base, foram inseridos textos racistas nelas. Contudo, para tornar o experimento mais realista foram utilizadas redações com temas que existem uma probabilidade maior de ter comentários preconceituosos [1].

Para cada redação, foram incluídas 3 sentenças com preconceito. É importante destacar que essas sentenças não estão contidas na base de dados extraída do twitter, para não causar problemas de *overfitting*. Com isso, o banco de dados ficou com 15 sentenças preconceituosas e 70 neutras. Abaixo segue um exemplo do trecho de uma redação retirada

[1] https://educacao.uol.com.br/bancoderedacoes/propostas/carta-convite-discutir-discriminacao-na-escola.htm

do Banco de Redações da UOL, sem sentenças preconceituosas.:

> "Lá eles são considerados heróis, aqui são apenas os bandidos da vez. Não se devem esquecer os casos de abuso de poder e de policiais corruptos, porém nem tão pouco colocar todos no mesmo patamar. Eles não são o grande inimigo! Mas, nada podem fazer sem a ajuda do governo, que deve reforçar as leis e criar novas de acordo com atual situação brasileira. O Brasil não deve ser o país da impunidade e, enquanto houver tanta desvalorização da classe, não haverá progresso."

Para o experimento foi inserido manualmente sentenças preconceituosas, como no exemplo:

> "Lá eles são considerados heróis, aqui são apenas os bandidos da vez. **Todo bandido é negro ou tem a alma negra,** não se devem esquecer os casos de abuso de poder e de policiais corruptos, porém nem tão pouco colocar todos no mesmo patamar. Eles não são o grande inimigo! Mas, nada podem fazer sem a ajuda do governo, que deve reforçar as leis e criar novas de acordo com atual situação brasileira. O Brasil não deve ser o país da impunidade e, enquanto houver tanta desvalorização da classe, não haverá progresso."

Das 15 frases preconceituosas inseridas nas redações 14 foram classificadas corretamente e apenas 1 não. Por outro lado, 15 sentenças que não deveriam ser marcadas como preconceituosas foram, deixando assim 56 sentenças neutras.

Para analisar esse resultados foram consultados dois professores que trabalham com correção de redações. A opinião deles é que a ferramenta é um ótimo auxílio para lidar com o problema relacionado a competência 5 do ENEM (respeitar os direitos humanos). Contudo, ela não deve ser utilizada para uma avaliação automática e sim para indicar para os professores as frases que tem indícios de preconceito, mas a decisão final continuaria com o avaliador. Ou seja, seria uma abordagem semi-automática.

5.3 Discussão dos experimentos

Os algoritmos que alcançaram os melhores resultados no geral foram os de Redes Bayseanas e SVM (SMO). Esse resultado já eram esperado visto que esses algoritmos tem se destacado na literatura de classficação de texto. Por outro lado, o RBF teve os piores resultados. Nos experimentos foi aplicado o método de validação cruzada divididos em dez pastas entre conjunto de testes e treinamento.

Em relação as características os grupos de características que tiveram melhor desempenho foram o MFD2, MFD3, Token-NGrams e Character-NGrams. O MFD2 obteve resultados maiores para os algoritmos SVM e árvore de decisão, enquanto o MFD3 alcançou maiores resultados para redes bayseanas e KNN.

É importante destacar que esses grupos de características englobam 2 ou 3 palavras de cada postagem e os termos pertencentes aos dicionários propostos, tratando-se assim de uma abordagem híbrida que utiliza dicionário e algoritmos de aprendizagem de máquina ao se tratar do MFD2 e MFD3.

A combinação dos algoritmos com o Character-NGrams apresentaram resultados interessantes, pois na maioria dos cenários o SVM apresentou os melhores resultados, porém no Character-NGrams e Token-NGrams os melhores resultados foram alcançados através das Redes Bayseanas, levando em consideração a frequência das palavras em português, consiederando unigramas e bigramas.

A tupla (algoritmo, características) que alcançou o melhor resultado foi o algoritmo Redes Bayseanas e o conjunto de característica Token-NGrams. Portanto, essa configuração foi utilizada para identificação de sentenças preconceituosas em textos de redações.

Capítulo 6

Considerações finais e Trabalhos Futuros

Este trabalho apresentou uma abordagem para identificação de textos com indícios de preconceitos para o idioma português. Na proposta foi utilizada uma abordagem híbrida combinado dicionários léxicos e algoritmos de aprendizagem de máquina.

Para responder a questão de pesquisa do trabalho a configuração utilizada nos experimentos apresentou resultados satisfatórios. Foi realizado um experimento para avaliar diferente combinações de características e algoritmos de aprendizagem de máquina para o problema citado. O melhor resultado encontrado combinou como características Token N-Grams. Os melhores algoritmos avaliados foram o SVM e Redes Bayseanas, atingindo uma f-*measure* de 83,60% e f-*measure* de 82,10%, respectivamente.

Como resultados alcançados, o presente trabalho foi aceito no Simpósio Brasileiro de Informática na Educação como artigo completo: Uma abordagem computacional de analise de opinião para identificação de preconceito em redações, além disso, foi desenvolvido um dicionário de preconceito em português brasileiro, podendo ser replicado para outras pesquisas.

Dentre as limitações dos trabalho, podemos destacar a adoção de uma estratégia de pesos dos dicionários léxicos, o desenvolvimento de um estudo qualitativo das palavras usadas, como também o desenvolvimento dos dicionários léxicos no idioma inglês para português, já que na presente abordagem foi utilizado o dicionário de Tulkens que é no idioma alemão, também é necessária uma avaliação quantitativa em redações, aumentar a quantidade de palavras nos dicionários e desenvolver uma interface para que possa ser utilizado por professores em sala de aula para correção de redações.

Como trabalhos futuros sugere-se: (i) a ampliação da base de dados para utilização de algoritmos de *deep learning*; (ii) obtenção de redações reais que contém sentenças com comentários preconceituosos, para uma melhor avaliação educacional; (iii) criação de uma ferramenta visual para que professores possam utilizar.

Afim de responder a questão de pesquisa foi desenvolvido este experimento identificando frases preconceituosas em redações com a combinação de dicionários léxicos e aprendizagem de máquina, o objetivo central foi alcançado sendo desenvolvido uma abordagem computacional híbrida.

Em relação aos objetivos específicos:

- **Identificar principais grupos de palavras com emoções com maior índice de preconceito**: foi verificado e selecionado na literatura um dicionário de Pasqualotti [34], que é composto de grupos de palavras com emoções

- **Criação de um banco de dados para identificação de preconceito em português**: foi desenvolvido um banco de palavras baseado no dicionário de Tulkens [43] na língua portuguesa.

- **Análise de diferentes algoritmos para identificação de frases com indício de preconceito em redações em português**: o trabalho avaliou cinco algoritmos para verificar o que apresentaria os melhores resultados.

- **Combinar técnicas de aprendizagem de máquina com dicionários léxicos para identificação de preconceito**: o principal resultado deste trabalho foi alcançado através da abordagem híbrida, a partir do desenvolvimento do dicionário léxico com termos abusivos e a avaliação dos algoritmos de aprendizagem de máquina.

Bibliografia

[1] Swati Agarwal and Ashish Sureka. But i did not mean it!âintent classification of racist posts on tumblr. In *Intelligence and Security Informatics Conference (EISIC), 2016 European*, pages 124–127. IEEE, 2016.

[2] Swati Agarwal and Ashish Sureka. Role of author personality traits for identifying intent based racist posts. In *Intelligence and Security Informatics Conference (EISIC), 2016 European*, pages 197–197. IEEE, 2016.

[3] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[4] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[5] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[6] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora, 2013.

[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora, 2013.

[8] Eliana Cristina Nogueira Barion and Decio Lago. Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, 3(3):123–140, 2015.

[9] Karin Becker and Diego Tumitan. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. *Simpósio Brasileiro de Banco de Dados*, 2013.

[10] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F Smeaton. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, pages 231–236. IEEE, 2009.

[11] Elmano Ramalho Cavalcanti, Elmano Pontes Cavalcanti, Carlos Eduardo Pires, Rodrigo Alves Costa, and Caroline Ramalho Cavalcanti. Detecção e avaliação de cola em provas escolares utilizando mineração de texto: um estudo de caso. *Revista Brasileira de Informática na Educação*, 19(02):56, 2011.

[12] Arjun Chaudhuri. *Emotion and reason in consumer behavior*. Routledge, 2006.

[13] Emanuel Coutinho, Leonardo Moreira, Gabriel Paillard, and Ernesto Trajano de Lima. Análise do sentimento de mensagens de chats em uma turma de graduação de um curso de educação à distância. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 5, page 1019, 2016.

[14] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[15] Mariza Miola Dosciatti, Lohann Paterno Coutinho Ferreira, and Emerson Cabrera Paraiso. Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. *ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil*, 2013.

[16] Rafael Ferreira Leite de Mello. Retriblog: um framework centrado na arquitetura para criação de blog crawlers. 2011.

[17] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[18] Edel Greevy and Alan F Smeaton. 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, 2004.

[19] A Redação Guia. Guia do participante. *Instituto Nacional de Estudos e. Pesquisas Educacionais Anísio Teixeira. Brasília*, 2012.

[20] INEP. Inep - instituto nacional de estudos e pesquisas educacionais| enem. DisponÃvel em: http://g1.globo.com/educacao/enem/2017/noticia/enem-2017-chega-a- 38-milhoes-de-inscritos-a-quatro-dias-do-fim-do-prazo.ghtml. Acesso em: 12 Julho 2017, 2017.

[21] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

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

[23] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–614. ACM, 2007.

[24] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 710:25, 2000.

[25] Mary McGlohon, Natalie S Glance, and Zach Reiter. Star quality: Aggregating reviews to rank products and merchants. In *ICWSM*, 2010.

[26] Edison Andrade Martins Morais and Ana Paula L Ambrósio. Mineraçao de textos. *Relatório Técnico–Instituto de Informática (UFG)*, 2007.

[27] Paula Camargo Nascimento. *DICIONÁRIO DE POLARIDADES PARA APOIO A ANÁLISE DE SENTIMENTO*. PhD thesis, Universidade Federal do Rio de Janeiro, 2014.

[28] Luís Filipe da Cruz Nassif. Técnicas de agrupamento de textos aplicadas à computação forense. 2012.

[29] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 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, 2016.

[30] John T. Nockleby. *Hate speech*. Leonard W. Levy, Kenneth L. Karst et al., 2000.

[31] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[32] Sabrina Panceri and Crediné de Menezes. Apoio a mediação pedagógica em um debate de teses utilizando técnicas de processamento de texto. In *Simpósio Brasileiro de Informática na Educação*, volume 26, page 977, 2015.

[33] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[34] Paulo Roberto Pasqualotti. Reconhecimento de expressões de emoções na interação mediada por computador. 2008.

[35] Roberto HW Pinheiro, George DC Cavalcanti, and Tsang Ing Ren. Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications*, 42(4):1941–1949, 2015.

[36] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.

[37] Solange O Rezende, Ricardo M Marcacini, and Maria F Moura. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, 7:7–21, 2011.

[38] Eldar Sadikov, Aditya G Parameswaran, Petros Venetis, et al. Blogs as predictors of movie success. In *ICWSM*, 2009.

[39] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural*

Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, pages 1–10, 2017.

[40] Malcon Anderson Tafner. Redes neurais artificiais: aprendizado e plasticidade. *Cérebro Mente, São Paulo*, 5, 1998.

[41] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Classification: basic concepts, decision trees, and model evaluation. *Introduction to data mining*, 1:145–205, 2006.

[42] Songbo Tan. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28(4):667–671, 2005.

[43] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. The automated detection of racist discourse in dutch social media. *Mirror*, 2015.

[44] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185, 2010.

[45] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680, 2015.

[46] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[47] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

[48] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[49] Leandro Krug Wives. *Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de"clustering*. PhD thesis, Universidade Federal do Rio Grande do Sul, 1999.

[50] Tae Yano and Noah A Smith. What's worthy of comment? content and comment volume in political blogs. In *ICWSM*, 2010.

Capítulo 7

Apêndice - dicionários

7.1 Dicionário abusivo - Tulkens

Abdel	Rahman	epíteto	algodão
Abdul	Rashid	vagabunda	toalha de cozinha
Abdullah	Samir	cadela	toalha
Abu Ahmad	Umar	bastardo	galinhas
Ahmed	walid	cadela	vestuário
Ashraf	yassine	pálido	fazendo barulho
Aziz	Youssef	véu	vacas
Bilal	marrom	burqa	coelhos
Fouad	negro	burka	rebanho
Hamed	menino preto	cobertor	rebanhos
Hamza	branco	dromedários	almofada
Hasan	cappuccino	extremista	pano
Hassan	chocolate	burro	lençol
Hicham	cacau	fiorentina	cordeiros
Hosseini	marrom escuro	flanela	galinha
Ibrahim	filho da puta	cabras	marroquino
Imam Karim	prostituta	filho da puta	marroquinos
Khaled	xícara de manteiga	filhos da puta cabra	mulas
Khalid	prostituta câncer	chão	mula
Khalil	cara	cigano	boné
Mohamed	livro para colorir	odeia a barba	Nápoles

carneiros	chimpanzé	lesma	lobo
carpete mergulhador tela	animal	lêndeas	vermes
lona	esquilo	ostra	traças
toalha	ouriço	ostras	javali
torino	javali	elefante	imigrante
porcos	furão	vermes	estrangeiro
carne de porco	camarão	orangotangos	ilegal
gado	cabra	enguia	imigrante
doente	abatidos	panda	imigração
búfalo de água	girafa	parasita	estrangeiro
envoltório	lagarto	parasitária	mais pobres
cigano	hamster	semear erro	mais necessitados
ciganos	caruncho	semear erros	as minorias étnicas
preto-saia	hiena	primatas	imigrantes
macaco	insetos	ratos	de forma diferente
piolhos	barata	predador	de outra forma
parasitas	baratas	carne	mais pobres
rato	gato	marisco	belga
macacos	galinha	fungos	especial
bactéria	vaca	leitão	estranho
babuíno	coelho	gafanhotos	alienígena
babuínos	caranguejo	avestruzes	cidadania
de cerveja	crocodilo	tigre	comunista
de origem animal	crocodilo	toxoplasmose	terceiro mundo
de insecto	galinha poedeira	intermédios hospedeiras	os países do terceiro mundo
de leitões	bosque piolhos	porco	peculiar
afídeos	formigas	pulgas	emigrante
pulgões	vira	tarântula	pior
ácaros	mexilhão	guaxinim	exportação
chimpanzé	inseto	vespas	falha

trabalhadores convidados	albanês	etíope	iemenita
importado	argelino	europeus	Camarões
louco	americano	do europeu	Camaroes
importação	angolano	Fiji	Cazaquistão
importações	Argentina	Filipino	queniano
as importações	Armênia	finlandês	Quirguizistão
em desvantagem	australiano	francês	coreano
com baixa escolaridade	Azerbaijão	da Geórgia	croata
logica	asiático	Gana	libanês
curiosamente	birmanês	grego	da Libéria
difícil	bolivianos	Guatemala	Líbia
desenvolvimento	bósnios	Guiné	luxemburguesa
países não-ocidentais	britânico	Guiana	macedônia
normais nacionais ilógicas em desenvolvimento	da bulgária	Haiti	Malásia
os países produtores ilegais	cambojano	Honduras	maltês
as remessas	canadense	húngara	marroquina
a dívida	centro africano	irlandês	marroquina
caiu	Chile	islandès	mexicano
especial	chinês	indiano	moçambicano
punitiva	cubano	indonésio	namibiano
terrorista	Chipre	iraquiano	Holandês
triste	dinamarquês	iraniano	nepalês
Autorizações de residência	Dominicana	israelita	da Nicarágua
irritante	alemão	Israel	o neozelandês
quarto mundo	equatoriana	israelenses	nigeriano
alienígena refugiados	egípcio	italiano	ao norte africano
aliens	inglês	Costa do Marfim	norte-coreano
patéticos	equatorial-guineeer	Jamaicana	norte africano
afegão	da Estónia	Japonesa	norueguês
africano	Etiópia	japonês	Uganda

ucraniana	branco-russa	estrangeiro	estados do Golfo
ucraniano	Zâmbia	jogador de fora	Gonzalo
austríaco	cigano	búlgaro	grego
austríacos	Zimbábue	cambojano	ortodoxo grego
paquistanesa	sul-africanos	canadense	Guiana
paquistanês	sul-coreano	catalã	haitiano
panamenhos	sueco	Chile	hebraico
paraguaios	afegão	Comunidade	Hindu
peruana	afegãos	congolês	hondurenhos
peruano	argelino	crioulo	húngaro
português	argelina	cubano	islandês
Romeno	requerente de asilo	manifestante	os islandeses
roma	assírio	Dominicana	indiano
russa	australiano	alemão	indonésio
Salvador	asiático	do Equador	indonésios
Arábia	padeiro	egípcio	indonésia
Eslovaca	bola	estónia	Interfederal
sudão	varanda	etíope	internacional
espanhol	Barbados	o parlamento euro	iraque
espanhola	Bangladesh	europeu	irã
sudanês	birmanês	europeus	os iraquianos
sírio	bispo	Fiji	islâmicos
Tanzânia	fazendeiro	filipino	israelita
tailandês	boliviano	finlandês	Israel
um tailandês	bornéu	franciscano	italiano
melodia ornamental	bosniano	francês	italianos
turco	de bósnia	batatas fritas	Costa do Marfim
uruguaio	Brasil	Génova	jamaicano
venezuelano	britânico	prisioneiro	japonês
Vietnamita	bruxelas	Gana	o Iémen

iemenita	mexicano	persa	Checoslováquia
iemenitas	moldova	pólo	chechenos
jesuíta	muçulmano	piscinas	ornamental sintonia
Camarões	extremista islâmico	polonês	turco
cantonês	terrorista islâmico	português	turcos
Cazaquistão	moçambicano	racistas	do Uruguai
Kenya	Namíbia	raça	venezuelano
Quênia	nepal	rebelde	música popular
quenianos	nepalês	o líder rebelde	galês
Quirguistão	Nicarágua	contabilidade	cidadão do mundo
curdo	nova Zelândia	romeno	música cigana
kuwait	novo grego	rússia	Muçulmanos
kuwaitianos	nigeriano	russo	islão
congolês	norueguês	escocês	judeu
cabeçalho	Uganda	escrivão	sunitas
coreano	ucraniano	senegal	Alá
croata	ucrânia	eslovaco	mensageiro
conterrâneo	coligação política oriental	Eslovénia	cristão
latina	timorenses	Eslováquia	cristãos
libanês	coligacao política oriental	Eslovaco	extremista
líbano	austríaco	sudanês	fanático
relógio	ortodoxos	somaliano	fascista
pulseiras	otomano	união soviética	o fundamentalismo
lituano	grego antigo	espanhol	fundamentalista
Luxemburgo	Paquistão	sírio	religiosa
Macaca	paquistaneses	sírio-ortodoxa	hindus
maltês	paquistanês	Taliban	infiel
mandarim	palestino	suspeito de terrorismo	Islamismo
Mediterrâneo	os palestinos	tailandês	o islamismo
égua	do Panamá	um tailandês	islâmico

Mourad	café	lenço de cabeça	iluminado
Mustaffa	castanho claro	vagão	luz
Mustaffa	mancha	Islam crítico	carro do cavalo
Nabil	preto	juventus	aves
Omar	rascista	gaze	roma
profeta	vermelho	perus	gado
rachid	castanho avermelhado	camelos	ovelhas
Mohammad	colorido	a barba de ódio	hipopótamo
Mohammed	lápis de cor	semeador de ódio	elefante

israelita	risco de incêndio	tipo
Jeremias	agressor	espécie
Isaías	fluência	refinado
judaico	pirralho	aparência
jesus	perigo de explosão	nojento
jihad	explosivo	supostamente
judeus	extremista	vírus
Alcorão	vazamento de gás	câncer da mama
marxista	agitadores	câncer do pulmão
Cristo	hipócritas	câncer do cólon
mulher islâmica	colapso	câncer da próstata
as mulheres muçulmanas	canibais	câncer da bexiga
nacionalista	covarde	câncer do pâncreas
não-judeu	esquisito	câncer do esófago
não-muçulmano	vândalo	câncer do rim
palestinos	vândalos	leucemia
populista	assassinato	câncer
Profeta	assassino	herpes
o Profeta	psicopata	gripe
profetas	assassino em série	vírus da gripe
rabinos	motorista escravo	bactérias
carne de porco	bajulador	HIV
devoto	radiológico	vírus
renegado	terrorista	
anti-sociais	tirano	
bandidos	sexo	
trapaceiro	raça	
carrasco	orelha raça	
água de extinção de incêndios	grupo raça	
combustível	puro sangue	

7.2 Dicionário neutro - Tulkens

minorias étnicas	nacionalista	Azerbaijão	Dominica
imigrantes	não ocidental	Ásia	República Dominicana
americano	nacional	Bahamas	dubai
requerente de asilo	remessas criminais	Tailândia	país escuro
nativo	supranacionais	Bangladesh	Alemanha
belga	terrorista	Bélgica	Equador
nacionalidade estrangeira	turco	Bósnia	Estado unitário
estrangeira nativa	autorização de residência	Botsuana	Egito
comunista	refugiado	Brasil	ilhas
política de defesa	estranho	Reino Unido	grupo de ilhas
diplomático	desconhecido	Grã-Bretanha	emirados
expatriado	marrom	Bulgária	inglaterra
automobilismo	preto	Cabo	Equador
europeu	vermelho	Camboja	Guiné Equatorial
expat	branco	Canadá	Estónia
trabalhadores convidados federais	Afeganistão	catedral	Etiópia
ilegal	África	ilhotas	Europa
é ilegal	africano	ceilão	europeu
imigrante	Alaska	pimentão	Fiji

Grã-Bretanha	Croácia	Coréia do Norte	Salalah
Guatemala	Líbano	ártico	salinas
Guiné	Líbia	noruega	Ilhas Salomão
Guiné-Bissau	Lituânia	oceania	salvador
Guiana	Luxemburgo	Uganda	Arábia
Haiti	Ilhas Virgens	Ucrânia	Arábia Saudita
Honduras	Macau	Uzbequistão	Arábia
Hong Kong	Macedónia	Omã	Arábia Saudita
Irlanda	madagascar	Ontário	Escócia
Islândia	Madura	Timor Leste	Senegal
india	maestra	Áustria	Sérvia
Indonésia	maghreb	Paquistão	Xangai
Irã	malawi	palestinos	Singapura
Iraque	Malásia	palestina	Eslováquia
Itália	Malta	Panamá	sudão
Costa do Marfim	Marrocos	Paraguai	Somália
Jamaica	méxico	Paramaribo	Tanzânia
Japão	monaco	paz	Tasmânia
Iêmen	Mongólia	peru	Tailândia
Jerusalém	Montenegro	Portugal	Tibete
Iugoslávia	Moçambique	Quatar	República Checa
Colônia do Cabo	nacional	Quebec	Tunísia
Cabo Verde	Namíbia	dicas de viagem	Turquia
Camarões	Holanda	república repúblicas	Uruguai
Cazaquistão	Nicarágua	rica	EUA
Quênia	Nova Zelândia	rocha	do Vaticano
kuwait	Nigéria	Roménia	cidade do Vaticano
congo	África do Norte	Rússia	Venezuela
coreia	Coréia do Norte	Ruanda	vicente
coreano	África do Norte	sacramento	vietnam

País de Gales	não-judeu	bosniano	Haiti
Yucatan	não-muçulmano	Brasil	Hindu
África do Sul	sunitas	britânico	húngaro
Coreia do Sul	sunita	bruxelas	irlandês
África do Sul	afegão	estrangeiro	islandeses
Coreia do Sul	África	Bulgária	islandês
Suécia	africano	da Bulgária	indiano
Suíça	africanos	birmanês	indonésio
o fundamentalismo extremista Comunista	albanês	Burundi	indonésios
fundamentalistas	albaneses	chileno	internacional
fanáticos	da Argélia	congolês	iraque
fundamentalistas	argelino	crioulo	iraquiano
religiosos	americano	cubano	irã
o Islã	amárico	dinamarquês	iraniano
muçulmanos	andorrano	demonstrador	islâmico
o islamismo	Andrey	República Dominicana	Israel
islâmico	angolano	alemão	italiano
israelita	árabe	egípcio	Costa do Marfim
jihad	aramaico	Inglês	jamaicano
judeus	argentina	etíope	japonês
califado	armênio	Fiji	Iémen
Alcorão	pobre rim	filipino	iemenita
marroquina	braços irlandês	finlandês	iemenitas
marxista	requerente de asilo	francês	Cabo verde
um muçulmano	australiano	batatas fritas	Camarões
islâmico	austrália	irlandês	cantonenses
muçulmano	australiano	Gana	Cazaquistão
mulher islâmica	belga	estados do Golfo	quenianas
as mulheres muçulmanas	bélgica	grego	queniano
nacionalista	bósnia	ortodoxo grego	música

nativo	América	clemente	Finlândia
indonésio	Andorra	colômbia	popular
internacional	Angola	congo	Fransisco
desfavorecidos	a antártica	cordilheira	Gabão
mal educado	antartica	Cornualha	Gâmbia
legal	arábia	costa	gaza
legalize	árabe	Cuba	Georgia
marroquino	ártico	Chipre	Gana
marroquinos	Argentina	Dinamarca	Granada
migrante	austrália	diáspora	Grécia
imigrantes	Argélia	china	Filipinas

curdo	novo grego
kuwait	norueguês
kuwaitianos	Uganda
congolês	Ucraniano
Coréia	ucrânia
coreano	o dinheiro do petróleo
croata	Omã
libanês	extremidades leste
líbano	austríaco
relógio	ortodoxo
pulseiras	otomano
libanês	grego clássico
líbio	paquistanês
cilindros	Paquistão
Luxemburgo	palestino
Malásia	os palestinos
da Malásia	Panamá
maltês	papiamento
Marrocos	oficial polícia
marroquino	piscina
em parte rústica	português
Mediterrâneo	sacerdote
mexicano	
extremistas muçulmanos	
terroristas muçulmanos	
Moçambique Moçambique	
o holandês	
holandês	
nepalês	
nova Zelândia	

7.3 Dicionário de emoções - Pasqualotti

decepcionado	desprezivel	desafortunado
derrotado	infeliz	desconsolado
desapontado	lamentavel	desgracado
frustrado	lastimoso	
abalado	miseravel	
abatido	patetico	
aflito	pobre	
agoniado	acanhado	
agonizante	constrangido	
angustiado	culpado	
angustiante	desonrado	
contrariado	encabulado	
estressado	envergonhado	
inquieto	humilhado	
perturbado	azarado	
preocupado	coitado	
triste	deploravel	

7.4 Lista de *stopwords*

- a, à, adeus, agora, aí, ainda, além, algo, algumas, ali, ano, anos, antes, ao, aos, apenas, apoio, após, aquela, aquelas, aquele, através;
- baixo, bastante, bem, boa, boas, bom, bons, breve;
- cá, cada, cedo, certamente, certeza, cima, coisa, com, como, contra, custa;
- da, dá, dão, daquela, daqueles, daquele, daquelas, dar, das, de, debaixo, demais, dentro, depois, desde, dessa;
- e, é, ela, ele, eles, em, embora, entre, era, és, essa, essas, esse, esses, esta, está, estão, estar, estas, estás;