Revista de Sistemas de Informação da FSMA
n. 23 (2019) pp. 2-9

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL ERI·GO

Classificação de Mensagens Em Língua Portuguesa Com Traços de Racismo No Twitter

Rodolfo Costa Cezar da Silva, Deborah Silva Alves Fernandes, Márcio Giovane Cunha Fernandes
Instituto de Informática – Universidade Federal de Goiás

***Resumo*—Com o desenvolvimento das redes sociais e Internet, usuários tem uma plataforma para expressar suas crenças e opiniões com maior facilidade e alcance. Algumas dessas opiniões podem ser de caráter pejorativo e discriminatório quanto as características de uma pessoa, como por exemplo a cor de sua pele. Este artigo descreve um processo de detecção de traços de racismo em mensagens em língua portuguesa no Twitter. Discute racismo em termos éticos e legais para a identificação conceitual, aborda análise de sentimento, suas possíveis abordagens e os possíveis níveis de análise. Caracteriza dados coletados quanto a frequência e localização. Ao final, discute-se os fatos percebidos durante o processo de coleta e classificação de dados.**

***Palavras-chave*—Análise de Sentimento, Racismo, Twitter, Aprendizado Supervisionado.**

Classifying Portuguese Messages With Traces of Racism on Twitter

***Abstract*—The development of social media and the Internet has given users a platform to express their beliefs and opinions with greater ease and outreach. Some of these opinions may be pejorative and discriminatory towards other users based only on characteristics such as the color of their skin. This article describe a process for the detection of traces of racism in Portuguese language messages on Twitter. It discusses racism in ethical and legal terms for conceptual definition, addresses Sentiment Analysis, its possible approaches and levels of analysis. It characterizes collected data regarding frequency and location. Finally, we discuss the facts perceived during the data collection and classification process.**

***Index Terms*—Sentiment Analysis, Racism, Twitter, Machine Learning.**

I. Introdução

COm o uso intenso e cotidiano das Redes Sociais Online (RSO) na atualidade, o Processamento de Linguagem Natural (PLN) possui como uma de suas aplicações notáveis a Análise de Sentimento (AS), cujo objetivo principal é automatizar a compreensão de opinião (o boca-boca na Internet) sobre algum tema no grande volume de dados não estruturados disponíveis eletronicamente na Internet[1].

Autor correspondente : Rodolfo Costa Cezar da Silva, *rodolfo-costa.ufg@gmail.com*

Opinião não é um fato objetivo, mas um conceito formado por diversas experiências vivenciadas por um indivíduo ou por um grupo[1]. A natureza dinâmica do conceito é percebida em cada relação estabelecida entre os entes sociais ou a partir de um fato novo observado. Tal comportamento dinâmico é potencializado quando o universo de dados e relações aumenta vertiginosamente, o que é observado com a inclusão das RSO no cotidiano mundial.

Não obstante, tal potencialidade ressalta aspectos positivos e negativos do processo cognitivo humano. Este artigo perpassa pelo interesse nos meios de automatização do reconhecimento daqueles negativos relacionados aos discursos de ódio e preconceito, entre os quais podemos incluir a AS.

Análise de Sentimentos, também conhecida por Mineração de Opinião, é a área de estudo que analisa opiniões, sentimentos, avaliações, apreços, atitudes e emoções das pessoas sobre entidades tais como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos [1]. Essa técnica tem sido amplamente utilizada em uma variedade de aplicações e em diversos nichos de pesquisa. Tais aplicações adotam a AS para fazer predições, auxiliar a tomada de decisão, classificar sentimento público, entre outras finalidades.

Em 2008, Pang e Lee [2] descreveram técnicas e abordagens que visam auxiliar sistemas de informação orientados a opinião. O foco da pesquisa era introduzir métodos que tratam os desafios apresentados pelas aplicações que envolvem AS e Mineração de Opinião tais como subjetividade e ambiguidade da linguagem natural, e trazer aplicações práticas para esses métodos.

Desde então, houve um grande desenvolvimento tanto no arcabouço teórico quanto prático, portanto avistamos a relevância e oportunidade de desenvolvermos e implementarmos estas ideias.

Desta feita, este artigo objetiva o desenvolvimento em primeira instância de elementos viabilizadores à caracterização de traços de racismo em mensagens textuais em língua portuguesa publicadas na RSO Twitter, assim como sua classificação utilizando abordagens supervisionadas bem conhecidas no âmbito de classificação automática de textos como a Regressão Logística e *Naive Bayes*, que serão avaliadas através de métricas como *F-score*,

Recall, *ROC*[1]. Para tanto, este artigo está organizado da seguinte maneira : na seção II são tratados racismo e injúria racial, além de exemplos de racismo na internet, especificamente em RSO. A seção III aborda conceitos de AS. A seção IV trata da estruturação, parâmetros e resultados do experimento realizado. Possíveis extensões e trabalhos futuros são descritos na seção V. Finalmente, as discussões e conclusões são descritas na seção VI.

II. Racismo

APESAR da população brasileira ser formada por aproximadamente 50.47% de pessoas pretas e pardas, segundo o Censo Demográfico do IBGE de 2010[2], o racismo ainda é uma questão presente no Brasil.

Embora seja um problema frequente, sua definição não é tão trivial, tendo em vista que ideias racistas podem ser expressas e percebidas de várias maneiras. Segundo [3], "racismo é o conjunto de teorias e crenças que estabelecem uma hierarquia entre as raças e etnias. É uma doutrina ou sistema político fundado sobre o direito de uma raça (considerada pura ou superior) de dominar as outras. Por fim, é um preconceito extremado contra indivíduos pertencentes a uma raça ou etnia diferente, considerada inferior".

Tratando-se de leis brasileiras, o crime de injúria racial está associado ao uso de palavras depreciativas referentes à raça ou cor com a intenção de ofender a honra da vítima. Já o crime de racismo, previsto na Lei nº 7.716/1989, implica em conduta discriminatória dirigida a um determinado grupo ou coletividade e, geralmente, refere-se a crimes mais amplos.

A. Racismo na Internet

O racismo existe muito antes da Internet e redes sociais, porém foi com o desenvolvimento destas que as pessoas passaram a ter uma plataforma para expressar e propagar suas opiniões, crenças e sentimentos com maior facilidade e visibilidade para outros usuários. Alguns casos podem ser citados:

> Em 2005, um estudante de Letras na Universidade de Brasília (UnB) discutia o sistema de cotas para negros na sua universidade pelo Orkut (rede social descontinuada em 2014). Durante essa discussão, o estudante se referiu aos negros e afrodescendentes como "burros", "urubus", "macacos subdesenvolvidos", entre outras ofensas [4].
>
> A jornalista Maria Júlia Coutinho, a Maju do "Jornal Nacional", foi vítima de comentários preconceituosos na página oficial do programa no *Facebook* (Figura 1), em julho de 2016, logo quando se destacou pela sua cobertura da previsão do tempo.
>
> No dia 30 de junho de 2018, um *youtuber* conhecido (Figura 2) fez comentários racistas durante a Copa do Mundo FIFA de 2018, sobre o jogador da seleção francesa, Kylian Mbappé, associando o jogador com suas possíveis habilidades de realizar arrastões na praia.

Fig. 1. Comentário racista na página do *Facebook* de Maju

Fig. 2. *Youtuber* faz comentário racista sobre jogador francês no Twitter.

Devido à subjetividade e ambiguidade, identificar traços de racismo automaticamente em textos é uma tarefa bastante desafiadora.

B. Trabalhos relacionados a Racismo nas Redes Sociais

Um sistema de classificação automática de textos com traços racistas é proposto em [5]. Treinaram uma *Support Vector Machine* (SVM) - vide seção III-B1 - com os padrões encontrados a partir de *Bag-of-Words*[3] (BOW) e bigramas. O corpus[4] de 3 milhões de palavras foi divido em *datasets* de diversos tamanhos que continham o mesmo número de documentos racistas e não-racistas. Utilizando um conjunto de treinamento de 2 mil documentos e outro de teste com 410 documentos, concluíram que a técnica de SVM combinada com BOW obteve melhor resultado quando comparada a técnica de SVM combinada com bigramas cujas taxas de precisão foram de 87.33% e 84.77%, respectivamente. A abordagem descrita no trabalho realizado em [6] tinha como finalidade a detecção automática de comentários racistas em redes sociais holandesas. Foram extraídos 5759 comentários de páginas de redes sociais que continham orientação racista. Esses dados foram classificados em três categorias (*"racist", "non-racist", "invalid"*) por dois anotadores, e um terceiro anotador que era acionado quando havia discordância entre os dois primeiros. Técnicas de dicionário LIWC (*Linguistic Inquiry and Word Count*) e SVM (*Support Vector Machine*) foram adotadas para classificação automática de comentários. Além do dicionário LIWC, foi usado um dicionário expandido

[1]ROC : acrônimo para *Receiver Operating Characteristic Curve*
[2]Disponível em : https://www.ibge.gov.br/
[3]A representação de Bag-of-Words mantém um vetor para cada documento/texto e cada valor no vetor é uma palavra associada a sua frequência em um dado documento ou texto.
[4]Corpus é uma coletânea ou conjunto de textos sobre determinado tema

que contém palavras relacionadas a discurso racista. Esse método obteve uma taxa *F-score* de 0.46 quando comparado com as anotações feitas manualmente.

Pesquisadores propuseram em [7] mapear e mensurar a ocorrência de *cyberbullying* contra professores no Twitter através de técnicas de aprendizado de máquina. Os dados foram coletados durante uma semana através da API do Twitter, pré-processados e classificados em três categorias: positivo, negativo e neutro. Para classificação utilizaram um classificador bayesiano (*Naive-Bayes*) que obteve uma acurácia de 87.1%.

Em [8] são apresentados modelos para classificação de discurso de ódio no Twitter que consideraram características como etnia, deficiência e orientação sexual. Utilizaram técnicas para extração de características dos textos e relação sintática e gramatical entre palavras. O artigo descrito explora a influência de características diferentes na tarefa de classificação. Com esta finalidade, são testadas várias combinações de características para verificar qual dessas leva a um melhor resultado. A combinação de bigramas até 4-gramas combinada com o gênero obteve os melhores resultados, com acurácia de 73.66%.

O artigo aqui descrito trata da classificação de mensagens em língua portuguesa com traços de racismo, algo inédito quando comparado ao artigos supracitados.

III. Análise de Sentimento

ANÁLISE de Sentimento, também chamada de Mineração de Opinião, é uma área que estuda as opiniões, sentimentos, avaliações, atitudes e emoções sobre entidades, tais como produtos, serviços, organizações, indivíduos, eventos, tópicos e seus atributos [9]. Uma opinião regular expressa o sentimento sobre apenas uma entidade ou aspecto [10], por exemplo : *"Essa geladeira é muito boa!!"*, enquanto uma opinião comparativa compara diversas entidades baseados nos aspectos que elas compartilham entre si, por exemplo : *"A imagem da TV Sony é melhor do que a Samsung"*.

A. Níveis de Análise

Em geral, a Análise de Sentimento trabalha em diferentes níveis de granularidade, e que pode ser feita em três niveis :

Nível de documento em que o trabalho consiste em classificar se um documento como um todo expressa um sentimento positivo ou negativo. Este nível de análise assume que cada documento expressa opiniões sobre uma única entidade.

Nível de sentença consiste em analisar a polaridade do sentimento de uma apenas sentença. Assume-se que em um documento pode conter várias sentenças que podem possuir um sentimento individual. Cabe ressaltar que postagens e comentários em mídias sociais como o Twitter seguem um padrão de sentenças curtas [11].

Nível de entidade e aspecto em que a granularidade é menor, tenta definir um sentimento sobre uma entidade, analisa diretamente a opinião em si, baseado na premissa de que uma opinião consiste em um *sentimento* e um *alvo*. O objetivo desse tipo de análise é descobrir as entidades e as respectivas opiniões sobre elas separadamente.

O alvo da opinião define a problemática da AS. A frase *"Embora o atendimento não seja bom, eu amo esse restaurante"* é um exemplo de uma sentença com sentimento geral positivo. Porém, assume valor negativo sendo alvo o "atendimento".

B. Abordagens

Nesta seção vamos ver diferentes formas de avaliar os sentimentos. Não é uma lista exaustiva, mas representa algumas das técnicas mais utilizadas na área.

1) Aprendizado de Máquina: O aprendizado de máquina é classificado em supervisionado, semi-supervisionado ou não supervisionado. O primeiro consiste em fornecer um conjunto de dados de treino previamente anotado manualmente ou por algum tipo de serviço de colaboração coletiva (*crowdsourcing*, do inglês), como o *Figure-Eight*[5], com alguma polaridade de sentimento a um classificador.

Exemplos de classificadores supervisionados são:

Support Vector Machines (SVM), cujo princípio consiste em determinar separadores lineares no espaço de busca, chamados de *hiperplanos*. Através dos dados de treinamento, a SVM cria um hiperplano que divide as classes.

Regressão Logística, modelo probabilístico que atribui probabilidades (B) para *features* (X), dado um conjunto de treino conhecido (y). Esse encontra o conjunto de probabilidades que maximiza a probabilidade de $P(XjB;y)$.

Naive-Bayes, que utiliza o *Teorema de Bayes* para determinar o conjunto de probabilidades.

Na abordagem não-supervisionada, também conhecida como aprendizado por observação e descoberta, não há o uso de dados de treinamento, nesse tipo de abordagem o classificador deve encontrar padrões automaticamente sem conhecimento prévio dos dados. Esse tipo de método geralmente usa abordagens léxicas para classificação que utilizam dicionários léxicos de sentimentos. Tais dicionários associam uma palavra com um significado quantitativo, que varia de $[\ \ 1; 1]$ de acordo com sua polaridade, ou qualitativo, que associam uma palavra a uma certa polaridade (positivo/negativo, feliz/triste).

Na abordagem semi-supervisionada, parte-se do pressuposto que o conjunto dos dados disponível para treinamento é formado por uma parte rotulada, em menor número, e outra não rotulada, em maior número [12]. A ideia é utilizar os exemplos rotulados para obter informações sobre o problema e utilizá-las para guiar o processo de aprendizado a partir dos exemplos não rotulados [13].

A escolha de qual abordagem utilizar depende se existe um conhecimento prévio dos dados ou não. Se existem dados sobre o problema que deseja-se resolver, provavelmente

[5]https://www.figure-eight.com/

a abordagem supervisionada é a mais recomendada e que trará os melhores resultados. Caso haja pouco ou nenhum conhecimento prévio sobre o problema, a abordagem mais indicada seria a abordagem semi-supervisionada ou a abordagem não-supervisionada.

2) Abordagem Baseada em Lexicon: Há a adoção de um lexicon para a AS que conta e atribui pesos a palavras relacionadas a sentimentos (*sentiment words*).

Umas das abordagens é a baseada em *corpus*, em que um conjunto base de *sentiment words* (palavras de sentimento), com polaridade conhecida, explora padrões de co-ocorrência para identificar novas palavras de sentimento e sua respectiva polaridade em um grande corpus, que pode ser previamente criado, ou utilizar serviços que disponibilizam *corpora*, como o WordNet[6], HowNet[7], entre outros.

IV. Experimento

A. Proposta de Arquitetura

Para atingir o objetivo proposto neste trabalho, faz-se necessária a execução de uma série de etapas consecutivas para que o experimento seja executado corretamente. A Figura 3 ilustra o modelo da arquitetura proposta para a realização do experimento de detecção de traços de racismo em *tweets*.

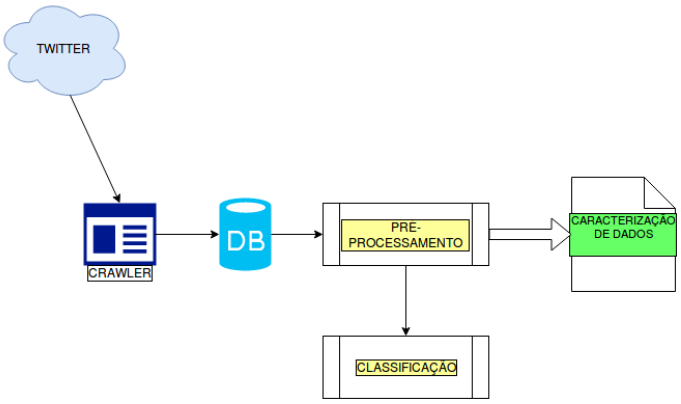

Fig. 3. Arquitetura proposta para análise de traços de racismo em tweets.

B. Coleta

A coleta de dados foi realizada entre os dias 04/06/2018 e 04/10/2018 no Twitter. A escolha dessa rede social ao invés de outras disponíveis se deu pelo fato de ser uma rede social popular no Brasil, e da plataforma oferecer uma *Application Programming Interface* (API) que permite, de forma fácil, coletar dados em tempo real.

O processo se inicia através do uso de um *crawler*, para coletar *tweets* da plataforma Twitter. À medida em que as mensagens são capturadas, são também armazenadas em um banco de dados. O *corpus* formado conta com 106.740 *tweets*.

[6]https://wordnet.princeton.edu
[7]http://www.keenage.com/

O escopo de coleta foi definido através de leituras e investigação de textos e casos sobre racismo na internet, redes sociais, jornais e palavras usadas em sistemas de buscas online. Após análises, optou-se também pela elaboração de um questionário para saber das pessoas quais palavras julgavam associadas ao racismo. Esse questionário foi disponibilizado durante 7 dias na conta de Twitter do pesquisador para que seus seguidores pudessem auxiliar a pesquisa. O questionário foi respondido por 16 pessoas e foi utilizado com uma ferramenta auxiliar para um melhor conhecimento do domínio do problema. Dentre as perguntas do questionário, destacam-se: "Você já sofreu por e/ou presenciou algum tipo de comentário que continha traços racistas em seu conteúdo?" e "O que caracterizava aquele texto como racista? Algum (ns) termo (s) em específico?".

Após as etapas descritas, as palavras selecionadas à composição do filtro de coleta de *tweets* foram: "senzala","gorila", "cabelo de bombril", "cabelo de esfregão", "nariz de nego", "nariz de nega", "tinha que ser preto", "tinha que ser preta", "preto da senzala", "preta da senzala", "preto da macumba", "nega macaca", "preto macaco", "preta macaca", "preta nojenta", "preto nojento", "criola", "crioula", "crioulo".

Um dos obstáculos encontrados na coleta é quando o nome do usuário contém algum termo do filtro. Por exemplo, um *tweet* enviado por um usuário chamado "spider gorila" foi coletado pelo *crawler*, mesmo que o corpo da mensagem não apresentasse comentário com traços racistas. Na seção IV-C é descrito como esse obstáculo foi contornado.

Outra dificuldade encontrada foi a lapidação do conjunto de termos que seriam buscados. Em coletas teste, os termos "nega" e "neguinha" estavam inclusos nos termos de busca, mas após uma análise visual dos resultados, percebeu-se que o primeiro termo na maioria da vezes, referia-se a uma conjugação do verbo "negar", e o segundo retornava várias mensagens de caráter apreciativo e afetuoso, e não discriminativo, portanto foram retirados do conjunto final de termos de busca.

C. Pré-processamento e rotulação

O pré-processamento realizado removeu *links*, pois observou-se que essas tratavam de mensagens que continham notícias que eram redirecionadas para outros sítios. Após removê-las, restaram 83900 *tweets* de interesse. Também foram removidos os nomes de usuários do corpo das mensagens, ou seja, palavras que iniciavam com '@', para a redução de ruídos antes de utilizar os dados em um classificador.

Baseado no método utilizado em [14], 2022 *tweets* foram rotulados por dois anotadores quanto à presença de traços de racismo nas mensagens. O conjunto de dados que foi apresentado aos anotadores continham *tweets* com textos distintos que foram apresentados de forma aleatória, em que cada anotador rotulava o *tweet* com 'sim', caso a mensagem contivesse traços de racismo, ou 'não', em caso

contrário. Do montante rotulado, 349 foram com 'sim' e 1676 com 'não'. A tabela I ilustra uma amostra dos dados anotados.

TABELA I
TABELA DE DADOS ROTULADOS

texto	traços de racismo?
(...)Não estamos mais, qual o problema? Seu preto macaco, Gorila	sim
Ah pronto, charuto preto da Macumba agora quer falar merda	sim
alguém corta a internet dessa escrava que fugiu da senzala	sim
Alguém já leu "casa grande e senzala"?	não
ainda não superei a morte da gorila koko	não
Agora sem está descursão de hoje este cara do boné preto	não

D. Nuvem de Palavras

Nuvem de palavras são utilizadas como uma maneira visual de apresentar a frequência de ocorrência de um certo termo ou palavra. Quanto maior o número de ocorrências de uma palavra, maior será a fonte utilizada para representá-la na nuvem de palavras. A figura 4 ilustra a nuvem de palavras do conjunto de dados em estudo.

Através de uma análise visual, constatou-se que dentre as palavras mais frequentes estão : "gorila", "crioulo" e "senzala". Geralmente, essas palavras são utilizadas como forma de desonrar pessoas negras através de comparação com animais, ou são palavras que fazem alusão ao período escravocrata.

Fig. 4. Nuvem de frequência de palavras do *dataset*.

É importante ressaltar que termos como "senzala" e "gorila" estão entre os termos mais frequentes, mas que podem ser utilizada de forma inócua, como pode ser visto na I. Isso mostra a dificuldade e a subjetividade do problema de classificação, pois a detecção do termo não infere o sentimento real da mensagem.

E. Geolocalização dos tweets

Dentre os conjunto de dados total, cerca de 64% (53589 *tweets*) continham informações geográficas disponibilizadas pelos usuários. Esses dados foram utilizados na construção de um mapa de calor apresentado na figura 5 que representa os locais com maior concentração de publicação durante o período de coleta no Brasil. É possível notar que a maior concentração de publicações é advinda das Regiões Sul e Sudeste, e o menor número em regiões como Norte e Nordeste do país. Isso pode ocorrer por inúmeras razões, como a taxa de acesso à internet, nível de analfabetismo, proporção de população negra nessas regiões, entre outras.

Segundo pesquisa realizada em [15], regiões como Norte e Nordeste enfrentam dificuldades para o acesso à internet devido a obstáculos como preço e disponibilidade, mas também destaca como obstáculo a falta de inclusão de indivíduos que não têm interesse e/ou não veem necessidade em acessar a rede.

O Censo Demográfico de 2010 do IBGE aponta que as regiões Sul e Sudeste tem população minoritariamente pretas e pardas, correspondendo a 20.6% e 46.6% de suas respectivas populações. De acordo com o mesmo Censo Demográfico, regiões Norte e Nordeste tem população majoritariamente preta e parda, correspondendo a 73.52% e 69%, respectivamente.

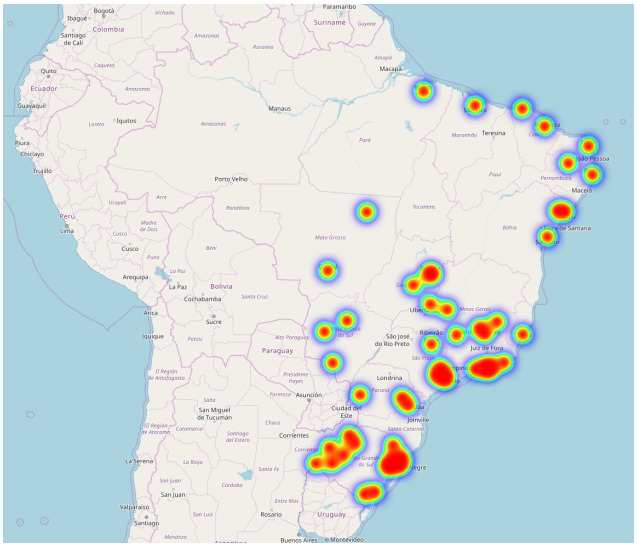

Fig. 5. Geolocalização dos *tweets*.

F. Classificação

No conjunto de dados de trabalho original existe o problema conhecido na literatura como o problema de classes desbalanceadas, que consiste em uma desproporção no número de dados de uma das classes do treinamento. Na problemática abordada temos essa desproporção nas classes de *tweets* que contém traços de racismo (349 *tweets*) e nas classes de *tweets* que não contém traços de racismo (1676 *tweets*). Segundo [16] , modelos que são criados sob tais condições têm tendências de serem modelos com alta acurácia global, porém triviais, que predizem quase sempre a classe majoritária e não caracteriza um modelo bem representativo.

Uma das soluções para tal problema de desbalanceamento é a utilização da técnica de balanceamento artificial, que consiste na remoção de instâncias da classe majoritária, a fim de construir um modelo mais robusto e mais preciso.

Modelos de classificação devem ser avaliados utilizando diversas métricas sensíveis a distribuição, e não apenas sua acurácia, pois um modelo de classificação pode ser bem preciso, porém não ser bem representativo, tendo em vista que as classes podem estar desbalanceadas. Por exemplo, um conjunto de dados de 100 instâncias, com 90 de uma classe x e 10 de uma classe y, modelado por um classificador pode ter uma acurácia de 90% se classificar todas as instâncias como classe x. O modelo hipotético descrito teria uma ótima acurácia, porém não seria bem representativo. Para uma análise completa e fiel de um modelo, deve-se avaliar métricas como *F-score, Recall e ROC*.

Para fins de análise e comparação, o montante total de *tweets* será particionado em dois *datasets* menores, cada um com suas características inerentes. O primeiro *dataset*, que será referenciado daqui em diante como **dataset1**, é um conjunto de dados em que foi aplicada a técnica de balanceamento artificial descrita anteriormente, além dos pré-processamentos descritos na sessão IV-C. O segundo *dataset* de estudo, que será referenciado daqui em diante como **dataset2**, é um conjunto de dados que tem as mesmas características do **dataset1**, com uma particularidade a mais que é a remoção de *stopwords*[8].

G. Experimento e resultados

O experimento proposto consiste em analisar dois classificadores diferentes sobre os dois conjuntos de dados. Nesse trabalho foram analisadas a acurácia, *F-score Recall* dos classificadores Regressão Logística (RL) e Naive Bayes (NB) utilizando o teste conhecido da literatura como *k-fold-cross-validation*[9] para $k = 10$, a fim de comparar como diferentes classificadores se comportam sob diferentes *datasets*. Os resultados do experimento são descritos nas tabelas II e III.

Apesar da remoção de *stopwords* ser amplamente utilizada no pré-processamento a priori de aplicar os dados em um classificador, por se tratar de termos que não carregam valor semântico relevante, existem autores como [17], [18] e [19], que afirmam que essas palavras de fato carregam informações relevantes e não devem ser removidas.

Neste trabalho foram utilizados *datasets* com e sem *stopwords* (*dataset1* e *dataset2*, respectivamente), a fim de mensurar a relevância desses termos, e percebeu-se resultados relativamente piores para um dos classificadores quando se removeram as *stopwords*.

Utilizando o classificador Regressão Logística, obteve-se acurácia de 76.336% e 73.798% para os *dataset1* e *dataset2*, respectivamente. Utilizando Naive-Bayes, onde alcançou acurácia de 66.176% e 66.577% para os *dataset1* e *dataset2*, respectivamente.

TABELA II
Resultados sobre o *dataset1*

	Acurácia	F-score	Recall
RL	76.336%	0.763	0.763
NB	66.176%	0.625	0.662

TABELA III
Resultados sobre o *dataset2*

	Acurácia	F-score	Recall
RL	73.798%	0.737	0.738
NB	66.577%	0.630	0.666

V. Trabalhos Futuros

TENDO em vista a dificuldade em classificar comentários/mensagens como "contendo traços de racismo" ou não, em trabalhos futuros pretende-se melhorar o processo de identificação de elementos textuais que podem configurar racismo, aplicando conhecimentos na área de ontologia para melhor delinear o domínio, assim como testar a aplicação de BabelNet[10], que é uma ferramenta que pode ser utilizada para criar uma rede semântica, expandir dicionários de sinônimos, etc.

Por se tratar de um tema severamente subjetivo e ambíguo, pretende-se melhorar a acurácia do classificador através do processo de rotulação, anotando uma maior quantidade de mensagens e adotando outros modelos de rotulação, incluindo características provenientes da estrutura do Twitter e *tweet* e dos próprios classificadores.

Em futuros trabalhos, almeja-se utilizar outros tipos de abordagens de AS, como por exemplo RNN (*Recurrent Neural Networks*), descrito em [20], a fim de comparar com os métodos propostos neste trabalho.

VI. Considerações Finais

FOI apresentado neste artigo um modelo de classificação de mensagens em língua portuguesa com traços de racismo na RSO Twitter. Para tal, foram coletadas 106.740 mensagens do Twitter que continham palavras de possível caráter racista que foram definidas através de estudos de casos de racismo, leituras, investigações de textos, jornais, e através de um questionário próprio. Após a coleta, foram aplicadas técnicas de pré-processamento com a finalidade de reduzir ruídos nas mensagens coletadas, que então passaram por um processo de rotulação por anotadores humanos quanto a presença de traços de racismo. Durante o processo de classificação, o conjunto de dados foi dividido em dois subconjuntos de dados menores que se diferenciavam apenas na presença ou não de *stopwords*. Nestes subconjuntos de dados foram aplicados os algoritmos de classificação Regressão Logística e *Naive-Bayes*, a fim de comparar os resultados obtidos.

[8]stopwords são palavras que carregam baixo valor semântico relevante para um classificador. *Stopwords* podem ser preposições, artigos, pronomes, etc.

[9]Também denominada de validação cruzada de k partes, consiste em dividir o conjunto todo em k subconjuntos mutuamente exclusivos do mesmo tamanho, onde $k - 1$ subconjuntos são utilizados para treino e 1 subconjunto para teste. Esse processo se repete k vezes alternando de forma circular o subconjunto de teste até que o modelo seja treinado e testado com todas as partes.

[10]https://babelnet.org/

Para o algoritmo de Regressão Logística, observou-se uma acurácia maior (aproximadamente 3% maior) quando aplicado sobre o subconjunto onde as *stopwords* não foram removidas, indicando a sua relevância semântica. Para o algoritmo *Naive-Bayes*, a diferença das acurácias para os dois subconjuntos de dados é ínfima (0.4%)

Durante o desenvolvimento desta pesquisa, foi possível perceber que o racismo é bem mais presente do que se imagina no Brasil. Por se tratar de um país fundado na miscigenação e diversidade, acredita-se que não existiriam práticas discriminatórias contra uma pessoa baseada em cor de pele, mas esse trabalho traz à tona um pouco da realidade de uma parte da sociedade brasileira que trata um problema social tão sério como vitimismo e/ou brincadeira.

Além de apresentar resultados satisfatórios, este trabalho proporcionou uma melhor compreensão do processo de tradução de um problema social, como o racismo, em uma modelagem computacional através do uso de aprendizado de máquina e AS, evidenciando as dificuldades encontradas e que devem ser exploradas ao tratar de problemas relacionados a linguagem natural, como: uso de gírias, abreviações, subjetividade e ambiguidade em textos, processo de rotulação, remoção ou não de *stopwords*, quais classificadores são mais apropriados para cada tipo de problema.

Referências

[1] B. Liu, *Sentiment Analysis and Opinion Mining*. 2012, p. 168, ISBN: 9781608458844.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends R in InformatioPang, B., & LeFoundations and Trends R in Information Retrieval, 1(2), 91–231.*, vol. 1, no. 2, pp. 91–231, 2006. DOI: 10.1561/1500000001. [Online]. Available: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.

[3] I. C. Martins, *O racismo nas redes sociais : O mundo virtual é feito por pessoas de carne e osso!* 2014. [Online]. Available: http://www.vvale.com.br/geral/racismo-redes-sociais/.

[4] ISTOÉ Independente, *O criminoso da internet*, 2015. [Online]. Available: https://istoe.com.br/434177_O+CRIMINOSO+DA+INTERNET/.

[5] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, 2004, ISBN: 1581138814. DOI: 10.1145/1008992.1009074.

[6] L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A Dictionary-based Approach to Racism Detection in Dutch Social Media," 2005.

[7] R. José De Alencar, "ESTUDO DA OCORRÊNCIA DE CYBERBULLYING CONTRA PROFESSORES NA REDE SOCIAL TWITTER POR MEIO DE UM ALGORITMO DE CLASSIFICAÇÃO BAYESIANO," pp. 5–1, 2012. [Online]. Available: http://periodicos.letras.ufmg.br/index.php/textolivre.

[8] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," pp. 88–93, 2016.

[9] G. Hirst and B. Liu, "SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES Sentiment Analysis and Opinion Mining Sentiment Analysis and Opinion Mining," 2012.

[10] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations," 2006.

[11] F. Benevenuto, "Métodos para Análise de Sentimentos em mídias sociais," 2015.

[12] B. Vicente, A. De Lima, V. P. Machado, R. De Melo, and S. Veras, "Abordagem Semi-supervisionada para Rotula ao de Dados,"

[13] R. Bruce, "A Bayesian Approach to Semi-Supervised Learning," 2001.

[14] G. Monteiro e Silva, "Aplicando Técnicas de Aprendizado de Máquina para Detecção Automática de Bullying no Twitter," 2017.

[15] "Panorama setorial da Internet Acesso à Internet no Brasil: Desafios para conectar toda a população," Tech. Rep., 2016. [Online]. Available: http://www.intgovforum . org / cms / IGF % 5C % 20Policy % 5C % 20Options % 5C % 20for % 5C % 20Connecting % 5C % 20the % 5C % 20Next % 5C % 20Billion % 5C % 20Compilation.pdf.

[16] M. C. Monard and G. E. A. P. A. Batista, "Learning with Skewed Class Distributions," 2003.

[17] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for Twitter sentiment analysis Conference or Workshop Item Alleviating Data Sparsity for Twitter Sentiment Analysis," pp. 2–9, 2012. [Online]. Available: http://www2012.wwwconference.org/.

[18] E. Martínez-Cámara, A. Montejo-Ráez, M. T. Martín-Valdivia, L. A. Ureña, and U. Ureña-López, "SINAI: Machine Learning and Emotion of the Crowd for Sentiment Analysis in Microblogs," Tech. Rep., 2013, pp. 402–407. [Online]. Available: http://www.cs.uic.edu/.

[19] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, 2013, ISBN: 9781450320351. DOI: 10.1145/2488388.2488442.

[20] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting Offensive Language in Tweets Using Deep Learning," 2018.

Rodolfo Costa Cezar da Silva Graduando em Ciência da Computação na UFG - Universidade Federal de Goiás. Durante a graduação foi monitor das disciplinas de Cálculo I e II. Participou do Programa Ciências Sem Fronteiras, onde cursou por 1 ano Ciência da Computação em Indiana University of Pennsylvania, localizada na cidade de Indiana, no estado da Pennsylvania, nos Estados Unidos.

Deborah Silva Alves Fernandes Possui graduação em Ciência da Computação pela Pontifícia Universidade Católica de Goiás (2001), mestrado em Engenharia Elétrica pela Universidade de Brasília (2006) e doutorado em Engenharia de Sistemas Eletrônicos e Automação pela UnB (2015). É professora adjunta da Universidade Federal de Goiás. Tem experiência na área de Ciência da Computação, com ênfase em Compiladores, programação de computadores e interface humano computador. Atua em pesquisas com análise de sentimentos, dinâmica humana e tomada de decisão com base em dados de redes sociais online.

Márcio Giovane Cunha Fernandes Profissional com primeira formação em Ciência da Computação pela PUC-Goiás (conclusão ano 1999). Cursou o mestrado em Engenharia Elétrica e de Computação na UFG-Goiás (conclusão 2003). A atividade profissional principal é a docência. Os principais interesses na computação e informática são: alfabetização de alunos em lógica de programação, desenvolvimento de software, computação social - análise de tendência em textos de microblogs . Durante o desenvolvimento profissional adquiriu experiências em coordenação de curso de bacharelado em sistemas de informação, em função de auditor de Certificado ISO, coordenação de TI e de equipe de desenvolvimento de software. É professor da Universidade Estadual de Goiás, curso de Sistemas de Informação - CCET. Doutorando em Engenharia de Sistemas Eletrônicos e Automação pela UnB, vinculado ao laboratório LARA. Atua em pesquisas com análise de sentimentos, dinâmica humana e detecção de eventos em RSO e textos disponíveis na internet.