Analisando Tweets Relacionados a Deficiências: uma Abordagem Baseada em Classificação

Ademir Baségio Junior
Universidade Federal do Oeste do Pará - UFOPA
Santarém, PA, Brasil
basegiojunior@gmail.com

Lucas Darlindo Freitas Rodrigues
Universidade Federal do Oeste do Pará - UFOPA
Santarém, PA, Brasil
lucas.darlindo@gmail.com

Antonio Fernando Lavareda Jacob Junior
Universidade Estadual do Maranhão - UEMA
São Luís, MA, Brasil
antonio.junior@professor.uema.br

Fábio Manoel França Lobato
Universidade Federal do Oeste do Pará - UFOPA
Santarém, PA, Brasil
fabio.lobato@ufopa.edu.br

ABSTRACT

Approximately 80 % of people with some form of physical, mental, or intellectual disability live in developing countries. These same countries have shown significant growth in the availability of the internet. Such facts reveal good possibilities regarding access to emotional support and experiences exchange among people with disabilities through social media. However, hate speech and derogatory comments about these people can be a recurring problem on these platforms. In order to identify these posts, this article features a classifier developed using *Twitter* posts related to disabilities. The results show that the tool developed is promising in detecting offensive and pejorative comments on this topic, which can be used in content management systems.

KEYWORDS

Discurso de Ódio, Mídias Sociais, Pessoas com Deficiência, Twitter

1 INTRODUÇÃO

Um estudo conduzido por [1] mostrou que cerca de um bilhão de pessoas possuem algum tipo de deficiência, correspondendo a aproximadamente 15% da população. Além disso, segundo [2], perto de 80% das pessoas com alguma deficiência física, mental ou intelectual moram em países em desenvolvimento. Uma pesquisa conduzida pelo *Pew Research Center* [3] mostrou um aumento das taxas de adultos usuários de internet pelo *smartphone* de 42% para 64% e do acesso de redes sociais em sites de 34% para 53% entre 2013 e 2018 nos países com esse mesmo grau de desenvolvimento, demonstrando que as redes sociais online e outras plataformas tem potêncial para contribuir no suporte à pessoas com deficiência e também em informar o público em geral.

De acordo com estudos, as redes sociais podem exercer um papel importante na vida de Pessoas Com Deficiência (PcD). Entre as diversas contribuições fornecidas pelas mídias sociais, [4] destaca que elas podem ajudar na propagação de informações úteis para diversos segmentos da população. E no trabalho de [5] é apresentado que as mesmas fornecem um canal de denúncia, induzindo a redes de apoio.

De acordo com o estudo de [6], pessoas com características comportamentais homofílicas, ou seja, aquelas com mais chances de se relacionar socialmente com pessoas semelhantes, também ocorrem em redes sociais online. Em redes sociais virtuais como o Twitter, diversos trabalhos dedicam-se a identificar comunidades que se agregam por homofilia [7]. Já em [8] é mostrado que interação entre pares na internet pode impactar na autoestima dos usuários, de modo a intensificar a sensação de pertencimento a um determinado grupo e incentiva essas pessoas a lidar com dificuldades e preocupações. Com base no crescente uso de redes sociais online anteriormente citado, essa se torna uma maneira propícia de usufruir dos benefícios proporcionados por relações e trocas de experiências entre PcD.

Contudo, casos de segregação e preconceito existem como em qualquer meio social e podem afetar negativamente a autoestima tanto de PcD quanto de seus familiares [9]. Nesses casos, frequentemente são usadas expressões características modificadas ao longo dos anos. Por exemplo, a palavra "retardado" evoluiu de um diagnóstico médico para fazer referência à um insulto [10]. A presença de discurso de ódio não é exclusivo de PcD, comentários racistas, homofóbicos e sexistas também são vastamente observados nas mídias sociais [11].

A partir desta problemática, este trabalho propõe o treinamento de um classificador para identificar de maneira automática qual o teor dos comentários realizados na rede social online *Twitter* à respeito das deficiências em geral. A hipótese é que os algoritmos *Term Frequency–Inverse Document Frequency* (*TF-IDF*) e *Bag-of-Words* (*BoW*), juntamente com os dados numéricos de engajamento são relevantes e podem contribuir com o resultado final. Dessa forma, *tweets* foram coletados e os classificadores *Random Forest*, *Support Vector Machine* (*SVM*) e *Naïve Bayes* (*NB*) foram treinados.

Desse modo, o restante deste trabalho está organizado da seguinte forma: na Seção 2 são apresentados os principais trabalhos relacionados a este. Na Seção 3 é descrita a metodologia adotada nos experimentos. Na Seção 4 os resultados obtidos e na Seção 5 as considerações finais e possíveis trabalhos futuros.

2 TRABALHOS RELACIONADOS

Toda rede social de suporte próxima de uma PcD tem uma importância significativa no bem-estar dela e dos seus cuidadores. Sendo assim, [12] realizou um estudo com 88 mães de crianças com deficiência de até 24 meses, analisando suas relações e o suporte social recebido por essas famílias. Tal estudo encontrou sinais de estresse, depressão e transtorno de ansiedade nas mães que tinham redes sociais de suporte consideradas fracas, e concluiu que redes sociais

para o apoio às mães com crianças com deficiência são de extrema importância para os mesmos e o seu meio familiar.

Na mesma linha de pensamento, [13] examinaram 200 estudantes universitários com deficiência e os tipos de contatos sociais que eles tinham. O principal objetivo foi entender qual o seu impacto no sucesso acadêmico desses estudantes em faculdades de 4 anos. Chegou-se à conclusão que o suporte social exerce diferença significativa no sucesso acadêmico desses estudantes e que diferentes tipos de suporte podem afetá-lo de maneiras distintas. O estudo também destaca a importância de programas de apoio que promovem conexões socialmente favoráveis a fim de atenuar efeitos negativos que afetam o desempenho acadêmico.

Ainda nesse sentido, estudos foram realizados no sentido de entender qual a influência do capital social na saúde mental e no desempenho em atividades fundamentais para a vida da PcD. Dessa forma, [14] entrevistou 15.028 adultos vivendo em residências privadas na Austrália e se baseou na conceptualização de Bourdieu a respeito do capital. Os resultados do estudo mostram que as maiores desigualdades entre pessoas com e sem deficiências ocorrem no suporte emocional. No entanto, a desigualdade de capital social ainda foi pequena se comparada à desigualdade em relação a auto avaliação de saúde.

Já no trabalho de [15], foram realizadas entrevistas com 18 jovens com deficiências físicas e 17 de seus familiares. O objetivo foi de avaliar as reflexões dos participantes a respeito dos benefícios de intervenções com objetivo de estimular o uso de internet para fins de participação social desses jovens. A conclusão foi de que estudos realizados com internet precisam de abordagens mais críticas para investigar as necessidades dos usuários e as barreiras no uso da tecnologia da informação dentro de subgrupos como esse.

Baseado no cenário apresentado, considera-se identificar textos em redes sociais tanto à um caráter positivo, no contexto de deficiências, quanto negativo, ou seja, ofensivo. Nessa linha de pensamento, [16] fez um estudo à respeito do uso dos classificadores *Support Vector Machine*, *NB* e *Maximum Entropy* na detecção da polaridade de *tweets* de uma agência gerenciadora de tráfego, obtendo 99% de acurácia. Já [17], propõe uma combinação de modelagem de tópicos baseada no *LDA* (*Latent Dirichlet Allocation*) juntamente com *Words Embedding* para obter baixa dimensionalidade aplicada à tarefa de classificação. O estudo obteve resultados significativos em comparação com o estado da arte.

No trabalho de [18] foram utilizadas diversas abordagens para avaliar a classificação multirrótulo de notícias por assunto. Para isso, foram usados os algoritmos *K- Nearest Neighbors*, *SVM* e Árvore de Decisão combinados com os modelos de representação de documentos *TF-IDF*, *Word2Vec* e o modelo próprio do autor *W2VP-IDF*. Os resultados foram positivos com a melhor performance sendo obtida pelo classificador *SVM* juntamente com o *TF-IDF*.

No contexto de análise de sentimentos, [19] realiza uma comparação entre algoritmos de aprendizado de máquina e abordagens léxicas. O estudo analisou mensagens capturadas do *Twitter* que tinham relação com as Olimpíadas de 2016 realizada no Rio de Janeiro. A conclusão foi que os algoritmos de aprendizado de máquina se saíram melhores que as abordagens léxicas, com destaque para o *SVM*, que obteve 89,5% de acurácia. Em [11] os autores aplicaram modelagem de tópicos, análise de sentimentos e classificação de conteúdo para comentários de notícias da plataforma G1, as notícias eram relacionadas a temática de deficiência.

[20] avaliou o uso do modelo *Word2Vec* para a análise da polaridade de sentimentos. O estudo foi realizando em textos curtos extraídos do *Twitter* na língua inglesa utilizando *NB* e *SVM* como técnicas de classificação. Já o presente trabalho trata de textos curtos no português brasileiro. Os resultados demonstram que o *Word2Vec* é uma ferramenta promissora, apesar dela necessitar de uma quantidade de dados maior do que aquelas que se pretende anotar para serem alcançados resultados eficazes.

[21] avaliou em experimentos extensivos vários classificadores de aprendizado de máquina, a saber, KNN, Gaussian Naïve Bayes, (SVM), Multinomial Naïve Bayes e Random Forest (RF) para a classificação de conteúdo de publicidade. Os autores utilizaram dados do Facebook e concluíram que, para o domínio estudado, o SVM apresentou resultados superiores do que os outros métodos testados.

3 METODOLOGIA

Nas subseções abaixo são apresentados os métodos utilizados para a captura e estudo dos *tweets*. Especificamente, são descritos os processos de seleção dos termos de busca, captura dos dados, do sistema de anotação utilizado, cálculo do índice de concordância entre os juízes, e a classificação dos *tweets* de acordo com as classes propostas.

3.1 Definição das Categorias e dos Termos de Busca

A definição das categorias para a classificação dos *tweets* foram definidas com base em [9]. As classes propostas estão dispostas conforme a Tabela 1, sendo que na coluna da esquerda estão as categorias e na direta está a descrição de cada uma delas.

No mesmo trabalho, os termos usados para realizar as buscas de *tweets* relacionados às deficiências foram delimitados em uma reunião com um psicólogo, um profissional da área de comunicação e um analista de redes sociais. No resultado da anotação realizada, 65,1% dos *tweets* correspondiam a classe "Outros". Resultado esse, que dificultou conclusões mais precisas por limitar a quantidade de dados relacionados as classes de interesse. Assim, para evitar o excesso de ruído nos dados, os termos mais ruidosos foram excluídos da busca.

Esses termos estão presentes na Tabela 2. Na coluna esquerda (termos excluídos das buscas), estão presentes todos os termos mais ruidosos, ou seja, que mais contribuíram para a captura de *tweets* da categoria "Outros". Na coluna da direita (termos usados nas buscas) estão os termos finais escolhidos para realizar a coleta dos dados usados neste trabalho.

3.2 Coleta de Dados

Os dados utilizados foram coletados exclusivamente da rede social online *Twitter*, utilizando a *API* fornecida pela própria plataforma, a *Streaming API*. O período de coleta foi distribuído entre os 6 últimos meses do ano de 2018, evitando-se datas importantes para o tema, o que poderia gerar enviesamento na amostra de dados coletados. Foi realizada uma atualização das características de cada um dos *tweets* ao final da coleta por causa do longo período.

Tabela 1: Categorias definidas para a classificação dos tweets.

Categorias	Descrição
Infomativo	*Tweets* com o objetivo de transmitir informações sobre o tema deficiência.
Ofensivo e Pejorativo	*Tweets* com que contêm termos com o fim de ofender alguém ou um grupo de pessoas.
Indignação e Denúncia	Se refere aos textos que fazem uma denúncia de crimes ou relatam uma indignação à algum caso ocorrido.
Relato de Experiências	Diz respeito a um acontecimento com o próprio autor da postagem.
Outros	Categoria destinada à *tweets* que não se encaixam nas outras.

Tabela 2: Termos excluídos (originalmente incluídos em [9]) e os termos usados na coleta de dados nesta pesquisa.

Termos excluídos das buscas	Termos usados nas buscas
Deficiência	Deficiência Visual
Deficiência Mental	Pessoa com deficiência
Cego(s)	Paralisia Cerebral
Cegueira	Lesão Medular
Surdo(s)	Espinha Bífida
Surdez	Mielomeningocele
Autismo	Baixa Visão
Autista(s)	Deficiência Auditiva
	Deficiência Física
	Deficiência Intelectual
	Amputação
	Síndrome de Down

As características consideradas relevantes para a classificação são mostradas na Tabela 3. A primeira característica (*tweet*) são os textos retornados pela *Streaming API* sem nenhum modificação. As características *followers*, *friends*, *retweeted_count* e *favorite_count* estão em formato numérico e também fornecidas pela *API*.

3.3 Sistema de Anotação e Validação

Com o objetivo de facilitar e agilizar a anotação dos dados, foi criado um sistema para que os juízes realizassem a mesma através dele. Ele tem interface simples, com login e botões de rádio para classificar o *tweet*. Pode ser executado em qualquer navegador, tanto em *smartphone* quanto em *desktops*. Pela simplicidade e rapidez de transição entre os *tweets*, é possível anotar uma quantidade significativamente maior de dados do que ao usar técnicas manuais como planilhas.

O projeto foi desenvolvido na linguagem de programação *PHP* juntamente com o *CodeIgniter Framework*, contando ainda com elementos em *Hypertext Markup Language* (*HTML*), *Cascading Style Sheets* (*CSS*) e *Bootstrap*. O padrão *Model-View-Controller* (*MVC*) foi usado e vem previamente acoplado com a estrutura de desenvolvimento do *CodeIgniter*, gerando alguma proteção e segurança mínima entre o usuário e o banco de dados não relacional *MongoDB*, onde ficam armazenados os dados coletados.

Na Figura 1 é possível visualizar a tela de anotação do sistema. Um contador mostra o número de *tweets* já classificados e quantos ainda faltam. Logo abaixo está o *tweet* ("Exemplo de um tweet."), as cinco opções para classificação pré-definidas neste trabalho estão logo abaixo junto com o botão de enviar. Após isso o usuário será redirecionado para o próximo *tweet*.

3.3.1 Índice de Concordância. Uma parte importante dos métodos de classificação supervisionada é que seja mantido um determinado nível de consistência nas classes alvo dadas para o método realizar o processamento. Para avaliar a concordância entre os três juízes com a tarefa de anotar o conjunto de dados, faz-se necessário técnicas que possam simplificar a visualização da validade do conjunto para classificação.

A medida estatística de [22] é utilizada para a valiar a confiabilidade da concordância entre um número fixo de avaliadores. Sua principal diferença em relação à medida mais usual para se avaliar o nível de concordância entre avaliadores, o *Kappa* de *Cohen*, é a possibilidade de ser utilizado em mais de dois avaliadores simultaneamente. Dessa forma, optou-se pelo uso da mesma, afim de comparar tanto a concordância entre os três juízes, quando a concordância entre pares de juízes, sendo possível assim identificar possíveis causas para baixos níveis de acurácia. Essa estratégia também foi adotada por [11, 21].

3.4 Pré-processamento

Para melhorar a performance do processo de classificação, foram usadas diversas técnicas de pré-processamento. Neste conjunto de dados em específico, após uma exploração prévia dos mesmos, notou-se a necessidade da retirada de elementos que poderiam influenciar no resultado final da classificação por se tratarem de textos muito informais.

No trabalho de [23], uma revisão da literatura analisou as principais técnicas de pré-processamento utilizadas na análise de sentimentos no português brasileiro. Seus achados mostram que algumas das técnicas que estão entre as com maior frequência de uso são: *stopwords*, *hashtag*, tokenização, menções, *URL*, pontuação, *steamming*, acentuação, entre outras. Dessa forma, foram adotados os seguintes processos:

- **Somente minúsculas**: Nesta etapa todas as letras são convertidas para minúsculas a fim de evitar que duas palavras que são iguais sejam consideradas diferentes, como por exemplo as palavras "Visão" e "visão";
- **Acentuação**: Nesta etapa todos os acentos são removidos também para evitar palavras iguais serem consideradas diferentes, como por exemplo: "Visão" e "Visao";

Tabela 3: Características do *tweet* e do usuário.

Id	Característica
tweet	Tweet
followers	Quantidade de usuários que seguem o autor do tweet
friends	Quantidade de usuários que o autor do tweet segue
retweeted_count	Quantidade de retweets da postagem
favorite_count	Quantidade de *likes* da postagem
fol_per_fri	Número de seguidores dividido pelo número de seguidos
fri_per_fol	Número de seguidos dividido pelo número de seguidores
n_urls	Número de *URLs*
n_mentions	Número de vezes que o *tweet* menciona outro usuário

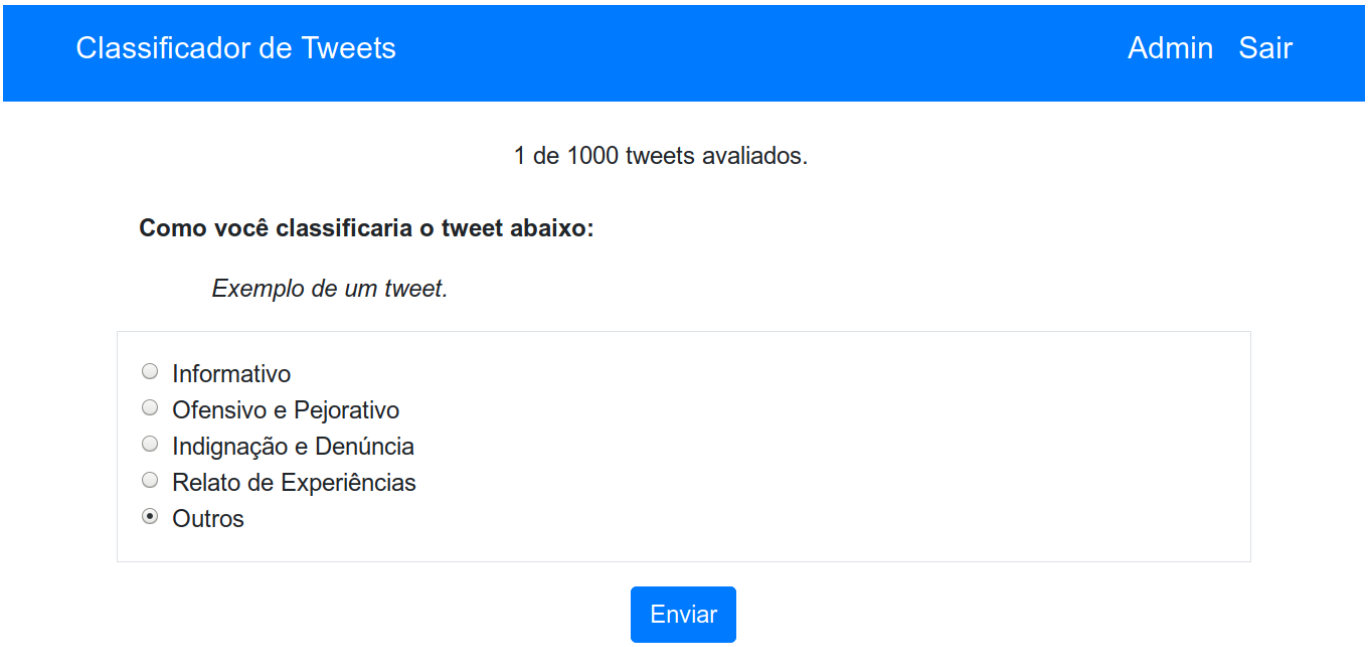

Figura 1: Interface do sistema criado para agilizar a anotação dos dados.

- **Pontuações e caracteres especiais**: Todos as pontuações e caracteres especiais são removidos para reduzir os possíveis ruídos no modelo;
- **Números**: Todos os números são removidos;
- ***Stopwords***: São removidas as palavras consideradas irrelevantes para o contexto do processamento de linguagem natural, como por exemplo: as, o, de, para;
- **Espaços duplicados e quebras de linha**: Para evitar que ocorra algum erro nos algoritmos utilizados, espaços duplicados e quebras de linha são eliminadas;
- ***Links***: Qualquer *link* para outro site é eliminado do texto também para reduzir ruídos;
- **Menções de usuários**: São removidas as menções de usuários dentro de *tweets* e um campo com a quantidade de menções de usuários presentes em um determinado *tweet* é adicionado;
- ***Stemming***: Nesta etapa palavras são reduzidas a sua raíz, como por exemplo as palavras "conectar", "conectando" e "conectado" seriam transformadas em "conect".

3.5 Treinamento do Classificador

Antes de ser feito o treinamento dos classificadores, os dados considerados importantes citados anteriormente foram salvos para o formato de arquivo *CSV*. Novas *features* foram identificadas a partir dos *tweets* coletados, são elas: quantidade de menções de usuário (*n_mentions*), quantidade de *URLs* (*n_urls*), divisão do numero de seguidores pelo número de seguidos (*fol_per_fri*) e divisão do número de seguidos pelo número de seguidores (*fril_per_fol*).

3.5.1 Representação de Texto como Vetor. Para ser possível realizar a classificação dos *tweets* coletados, faz-se necessária a utilização de técnicas que possam preparar os mesmos para o processamento. Uma das duas técnicas utilizadas foi o *TF-IDF*, que em tradução

livre significa frequência do termo–inverso da frequência nos documentos. O *TF* é responsável por dizer qual a importância das palavras em relação a todo o texto por meio da frequência que essa palavra aparece no texto. Já o *IDF*, tenta equilibrar o calculo final para casos onde palavras comuns que aparecem muito frequentemente mas tem pouca ou nenhuma importância, como a palavra "a", não afetem negativamente o a acurácia do modelo. Para isso, ele adiciona o inverso da frequência na equação, reduzindo o peso dessas palavras. Outra abordagem explorada foi a *BoW*. A mesma tem funcionamento parecido com a técnica anterior, contando quantas vezes uma palavra aparece no documento mas sem dar peso a elas.

3.5.2 Algoritmos de Classificação e os Conjuntos de Dados. Em relação aos métodos de classificação, uma metodologia muito parecida com a utilizada em [24] foi adotada. Foram escolhidos os classificadores *NB*, *Random Forest* e (*SVM*) juntamente com o *K-fold* com *10-folds*. Partindo disso, alguns *datasets* foram organizados e comparados nos resultados:

- ***df1***: Neste *dataset* constam todos os dados fornecidos diretamente pela *API* do *Twitter* juntamente com as características como número de menções de usuário e número de *URLs*, assim como consta na Tabela 3;
- ***df2***: Este *dataset* é composto pela matriz gerada pelo algoritmo *TF-IDF*;
- ***df3***: Este *dataset* é composto pela matriz gerada pelo algoritmo *BoW*;
- ***df1-df2***: Este *dataset* é composto pelas características da Tabela 3 e pela matriz gerada pelo algoritmo *TF-IDF*;
- ***df1-df3***: Este *dataset* é composto pelas características da Tabela 3 e pela matriz gerada pelo algoritmo *BoW*;

4 RESULTADOS

A análise dos dados coletados conforme apresentado na metodologia deste trabalho mostrou baixa eficácia dos algoritmo utilizados na tarefa de classificação dos *tweets*. Esses resultados são explorados nas seções a seguir.

4.1 Coleta e Anotação dos Dados

Foram coletados um total de 21.420 *tweets* válidos, isto é, desconsiderando aqueles repetidos, provenientes somente do português. As buscas foram realizadas tomando como base os termos previamente definidos neste trabalho com o uso da *Streaming API* fornecida pela plataforma *Twitter*, de onde foram realizadas as coletas. O período de coleta foi distribuído pelos 6 últimos meses de 2018, evitando-se datas representativas que diziam respeito ao tema para reduzir a possibilidade de enviesamento dos dados.

Do total de dados coletados, foram separados 1.431 *tweets* para ser realizada a classificação. Para garantir uma precisão mínima na representatividade da amostra, foi definido um intervalo de confiança de 95% com uma taxa de erro de 5%. Realizando o calculo amostral, temos como resultado 378 *tweets* necessários para o erro mínimo definido. No entanto, a agilidade do sistema de anotação permitiu a expansão considerável desse número para 1.431 *tweets* selecionados aleatoriamente. Isso garantiu uma margem de erro de 2,5%.

Na Figura 2, pode-se ver o resultado da distribuição dos *tweets* classificados considerando as ocasiões em que dois ou mais juízes estavam de acordo quanto à classe. Comparando com o trabalho [9], é perceptível a redução dos ruídos da classe "Outros" de 65,1% para 24,74% e um aumento das classes "Informativo", "Relato de Experiências" e "Indignação e Denúncia", sendo que todas estavam abaixo de 10% no trabalho citado.

Dessa forma, foi possível uma melhor análise das classes de interesse. Por outro lado, ocorreu uma redução significativa não prevista da classe "Ofensivo e Pejorativo", o que acaba por prejudicar a análise da principal classe de interesse na pesquisa, que são os comentários pejorativos e discursos de ódio.

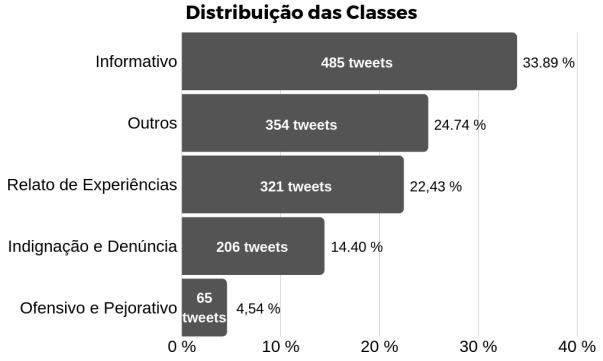

Figura 2: Distribuição dos *tweets* classificados entre as categorias.

Para dar validade ao estudo e às métricas utilizadas, viu-se necessário avaliar a concordância entre os três juízes. Para isso, foi feito o uso da técnica [22], afim de medir a concordância entre os avaliadores e também individualmente entre eles. A Figura 3 mostra os resultados do algoritmo para algumas combinações entre os juízes, observando-se um nível de concordância geral mediano na escala de 0 a 1. O juiz 3 foi o que destoou mais dos outros dois.

4.2 Classificação

Por meio da inspeção da Tabela 4 é possível conferir as medidas de acurácia, *average precision*, *average recall* e *average F1-Score*, obtidas a partir do treinamento dos classificadores em todos os cenários propostos. Cada algoritmo foi executado 10 vezes com o *Cross Validation* em cada um dos *datasets*. O *F1-Score* foi escolhido como principal medida de referência para comparar os classificadores por incluir tanto a precisão quando o *recall* no seu cálculo. Utilizando o mesmo como medida, o *NB* foi o melhor e esteve um pouco acima do *SVM* em todos os *datasets*. A exceção foi no *df1*, onde *SVM* e *Randon Forest* (*RF*) empataram.

Apesar disso, não houve diferença significativa de desempenho máximo alcançado entre o *SVM* e o *NB*. O melhor *F1-Score* alcançado pelo *NB* foi no *dataset* "*df3*" com 0,5; O *SVM* foi melhor no "*df1-df2*" conseguindo 0,48. Portanto, obtendo uma eficácia muito semelhante.

Também não ouve mudança considerável entre o *TF-IDF* e o *BoW*. Quando comparados os conjuntos "*df2*" e "*df3*" percebe-se uma diferença de somente 0.01 de *F1-Score* do melhor resultado do primeiro *dataset* (0,49) para o melhor resultado do segundo (0,50), ambos obtidos com o *NB*. Ainda se considerarmos os conjuntos

Tabela 4: Resultado da classificação de cada um dos métodos

Resultados					
Dataset	**Algoritmo**	**Acurácia**	**Avg. Prec.**	**Avg. Recall**	**Avg. F1**
df1	SVM	0,45	0,38	0,34	0,34
	RF	0,45	0,37	0,35	0,34
	NB	0,38	0,32	0,37	0,28
Dataset	**Algoritmo**	**Acurácia**	**Avg. Prec.**	**Avg. Recall**	**Avg. F1**
df2	SVM	0,53	0,60	0,45	0,47
	RF	0,52	0,56	0,40	0,40
	NB	0,56	0,50	0,49	0,49
Dataset	**Algoritmo**	**Acurácia**	**Avg. Prec.**	**Avg. Recall**	**Avg. F1**
df3	SVM	0,53	0,46	0,45	0,45
	RF	0,49	0,47	0,39	0,38
	NB	0,56	0,49	0,50	0,50
Dataset	**Algoritmo**	**Acurácia**	**Avg. Prec.**	**Avg. Recall**	**Avg. F1**
df1-df2	SVM	0,55	0,52	0,47	0,48
	RF	0,54	0,43	0,40	0,39
	NB	0,56	0,52	0,48	0,47
Dataset	**Algoritmo**	**Acurácia**	**Avg. Prec.**	**Avg. Recall**	**Avg. F1**
df1-df3	SVM	0,52	0,47	0,46	0,47
	RF	0,51	0,43	0,41	0,41
	NB	0,56	0,49	0,50	0,49

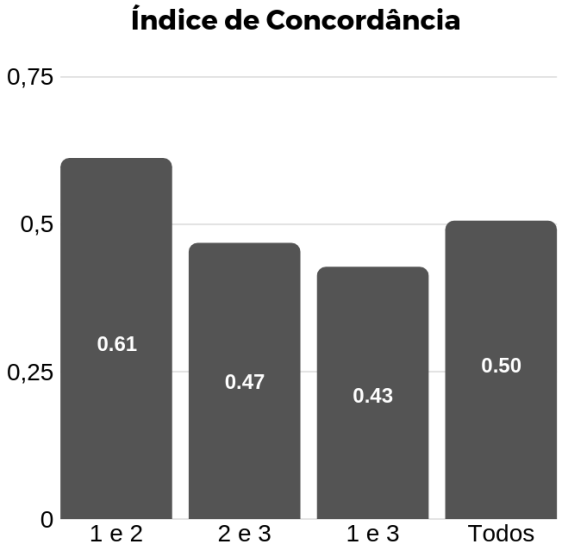

Figura 3: Índice de concordância entre os juízes 1, 2 e 3.

em que houve a combinação dos *datasets*, os melhores *F1-Score* dos modelos tiveram variação insignificante.

Em relação conjunto em que não houve processamento de texto (*df1*), onde os dados dizem respeito principalmente a dados de engajamento, o resultado ficou um pouco abaixo dos que houveram processamento de texto. O *F1-Score* variou de 0,28 no *NB* para 0,34 no *Random Forest* e *SVM*. Ao contrário do esperado, não se percebeu nenhuma alteração perceptível na acurácia quando combinado esse com os outros *datasets*.

Para entender melhor a distribuição de previsão entre as diversas classes propostas, o *dataset df3* com o classificador *NB* foi escolhido por ter obtido o melhor *F1-Score*. Conforme a matriz de confusão na Figura 4, é possível perceber que as classes "Informativo" e "Relato de Experiências" tiveram maior proporção de acertos, com 72% e 65% respectivamente. Nesse mesmo sentido, a classe que teve a menor proporção de acertos foi a "Ofensivo e Pejorativo", com 29%. Isso demonstra uma possível relação entre o desbalanceamento do número de amostras das classe com o resultado negativo obtido pelo classificador, já que a classe com menor número de amostras corresponde à classe com menor proporção de acertos. Isso evidencia a necessidade de reduzir os ruídos da classe "Outros" e balancear o número de *tweets* das classes alvo afim de obter o sucesso desejado.

4.3 Discussões

Para alimentar os classificadores do estudo, usaram-se dados captados a partir da rede social *Twitter* que tivessem algum termo incluídos naqueles definidos previamente. Técnicas de pré-processamento foram utilizadas para remover ruídos e melhorar a precisão do modelo final. Uma pesquisa bibliográfica apontou os principais trabalhos da área e quais algoritmos de aprendizado de máquina são utilizados mais frequentemente nesse tipo de análise.

A definição dos termos utilizados para a realização da coleta de dados foi feita baseado no trabalho de [9]. Nesse, houve uma porcentagem muito grande de *tweets* correspondentes a classe "Outros". Dessa forma, os termos atribuídos a essa classe foram retirados com o fim de reduzir os ruídos. No entanto, isso provocou um efeito inesperado da redução do número de ocorrências da classe "Ofensivo e Pejorativo". Esse desbalanceamento entre as classes e o número reduzido de exemplos utilizados para treinar os modelos pode ter reduzido significativamente a precisão do mesmo.

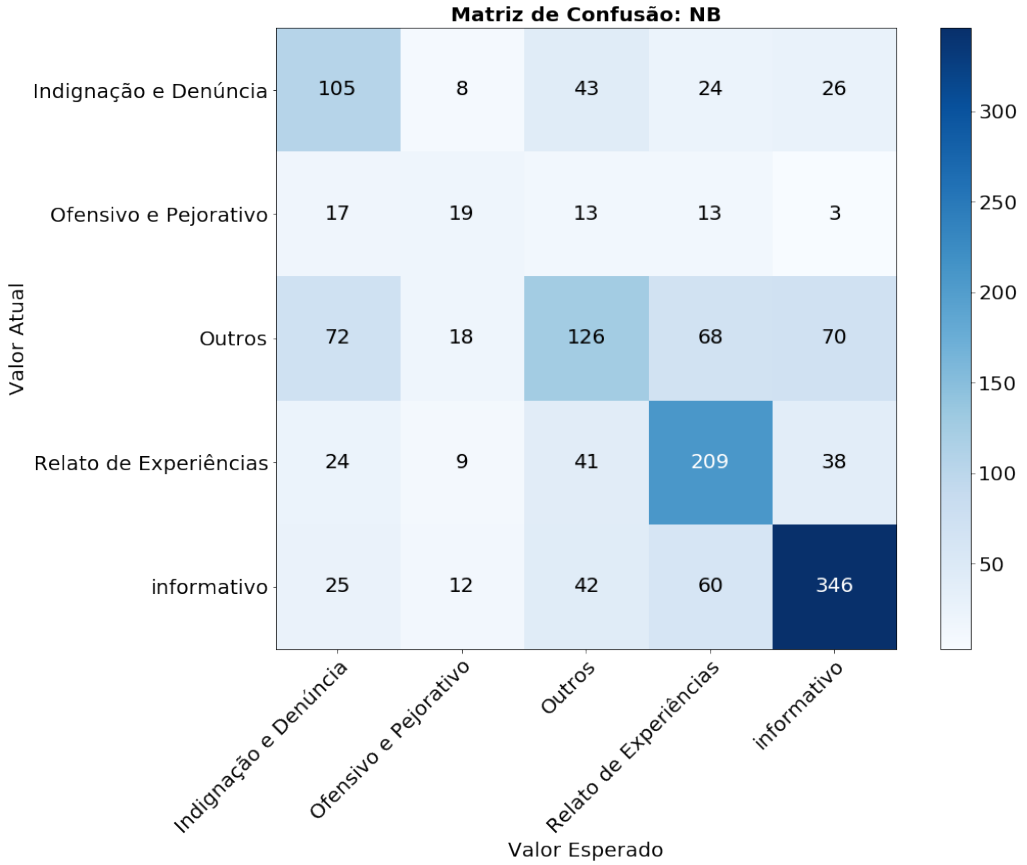

Figura 4: Matriz de confusão para o *dataset df3* com o algoritmo *NB*.

Isso é demonstrado claramente na matriz de confusão da Figura 4. A classe com menor número de *tweets* anotados, "Ofensivo e Pejorativo", foi a que claramente apresentou o pior desempenho entre todas. De 65 *tweets*, somente 19 foram atribuídos corretamente a essa classe.

Outra ameaça a validade do estudo é a medida estatística *Fleiss' kappa*. A mesma foi utilizada para dar validade ao modelo avaliando a concordância entre os três juízes no processo de anotação dos dados coletados. Os resultados dessa medida mostraram que houve uma concordância abaixo do desejável, o que pode acabar limitando a capacidade de previsão do classificador, visto que uma discordância muito alta entre os juízes indica muitas situações em que um mesmo *tweet* é atribuído a classes diferentes por diferentes juízes. A Figura 3 demonstra isso.

Todos os algoritmos utilizados para realizar a classificação e todos os *datasets* selecionados também apresentaram resultados abaixo do esperado. O *F1-Score* e as outras métricas utilizadas ficaram abaixo de 0,5 em quase todas as combinações de algoritmos e *datasets*. Dessa forma, não foi possível alcançar um modelo que possa conseguir prever as classes propostas com eficácia.

5 CONSIDERAÇÕES FINAIS

Neste trabalho foi abordado o treinamento de um classificador com objetivo de identificar o posicionamento dos usuários do *Twitter* no Brasil a respeito de deficiências. Para isso foram coletadas publicações relacionadas ao tema utilizando a *API* fornecida pela própria plataforma e baseando a escolha dos termos para a busca no trabalho de [9]. Um sistema para anotação de dados foi criado com o objetivo de aumentar o número de anotação possíveis por parte dos juízes. A partir desses dados coletados e anotados foi realizado o treinamento dos algoritmos *Random Forest*, *Support Vector Machine* e *Naive Bayes*.

Foram coletados um total de 21.420 *tweets* e desses, 1.431 foram anotados por três juízes. A avaliação da concordância entre eles apontou um nível médio de 0,5 em uma escala de 0 a 1, sendo que um dos juízes teve uma discordância maior em relação aos outros dois. Utilizando somente os dados onde dois ou mais avaliadores consentiram da resposta, foram treinados modelos nos algoritmos *Random Forest*, *SVM* e *NB* utilizando o *TF-IDF* e *BoW* para fazer a representação do texto como vetor.

Os resultados obtidos apontaram que as medidas de acurácia ficaram abaixo do desejado em todos os casos testados. Ao contrário do esperado que o *SVM* tivesse os melhores resultados, o *NB* foi o

que se saiu melhor, com um *F1-Score* chegando a 0,5 quando usado juntamente com o *NB* no *dataset df3*.

Como trabalhos futuros, propõe-se três possíveis abordagens para chegar mais próximo do objetivo deste trabalho. A primeira é a alteração nos termos usados na busca a fim de balancear a base de dados. Dessa forma, seria evitado o caso ocorrido com a classe "Ofensivo e Pejorativo", que obteve somente 65 *tweets* anotados. Novas maneiras de realizar representação do texto como vetor devem ser exploradas para tornar mais eficiente resultado da classificação. Por fim, um classificador poderia ser treinado com o objetivo exclusivo de identificar os *tweets* correspondentes à classe "Outros", reduzindo os ruídos no reconhecimento das classes alvo.

AGRADECIMENTOS

Os pesquisadores do Grupo de estudo e pesquisa em Computação Aplicada agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) no desenvolvimento deste trabalho, através da concessão de bolsa de pesquisa pelo Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI).

REFERÊNCIAS

[1] World Health Organization and others. Disability and health. *https://www.who.int/en/news-room/fact-sheets/detail/disability-and-health*, 2018.

[2] Ameneh Setareh Forouzan, Abolfazl Mahmoodi, Zahra Jorjoran Shushtari, Yahya Salimi, Homeira Sajjadi, and Zohreh Mahmoodi. Perceived social support among people with physical disability. *Iranian Red Crescent Medical Journal*, 15(8), 2013.

[3] Jacob Poushter, Caldwell Bishop, and Hanyu Chwe. Social media use continues to rise in developing countries but plateaus across developed ones. *Pew Research Center*, 22, 2018.

[4] Álvaro Figueira and Nuno Guimarães. Detecting journalistic relevance on social media: A two-case study using automatic surrogate features. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017.

[5] L. V. A. Caldas, A. F. L. Jacob, S. S. C. Silva, F. A. R. Pontes, and F. M. F. Lobato. Development of a social network for research support and individual well-being improvement. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug 2018.

[6] Michał Bojanowski and Rense Corten. Measuring segregation in social networks. *Social Networks*, 39, 2014.

[7] Wendel Silva, Ádamo Santana, Fábio Lobato, and Márcia Pinheiro. A Methodology for Community Detection in Twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 1006–1009, 2017. ISBN 978-1-4503-4951-2. doi: 10.1145/3106426.3117760. URL http://doi.acm.org/10.1145/3106426.3117760.

[8] JA Naslund, KA Aschbrenner, LA Marsch, and SJ Bartels. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2), 2016.

[9] Fábio Manoel França Lobato, Marcelo da Silva, Krislen Coelho, Simone da Costa Silva, and Fernando Pontes. Vamos falar sobre deficiência? uma análise dos tweets sobre este tema no brasil. In *7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*. SBC, 2018.

[10] Kathy McKay, Stuart Wark, Virginia Mapedzahama, Tinashe Dune, Saifur Rahman, and Catherine L Mac Phail. Sticks and stones: How words and language impact upon social inclusion. *Journal of Social Inclusion*, 6, 2015.

[11] Lucas D F Rodrigues, Jorge L. F. Da Silva Júnior, and Fábio M. F. Lobato. A culpa é dela! É isso o que dizem nos comentários das notıcias sobre a tentativa de feminicıdio de Elaine Caparroz. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 47–58. Sociedade Brasileira de Computação - SBC, jul 2019. doi: 10.5753/brasnam.2019.6547. URL https://sol.sbc.org.br/index.php/brasnam/article/view/6547.

[12] Janice E Kilburn and Cheri J Shapiro. The structure and function of social networks of mothers of young children with disabilities. *Topics in Early Childhood Special Education*, 2018.

[13] Allison Lombardi, Christopher Murray, and Jennifer Kowitt. Social support and academic success for college students with disabilities: Do relationship types matter? *Journal of Vocational Rehabilitation*, 44(1), 2016.

[14] Johanna Mithen, Zoe Aitken, Anna Ziersch, and Anne M Kavanagh. Inequalities in social capital and health between people with and without disabilities. *Social Science & Medicine*, 126, 2015.

[15] Lareen Newman, Kathryn Browne-Yung, Parimala Raghavendra, Denise Wood, and Emma Grace. Applying a critical approach to investigate barriers to digital inclusion and online social networking among young people with disabilities. *Information Systems Journal*, 27(5), 2017.

[16] Clarissa Castellã Xavier. Polarity classification of traffic related tweets. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2018.

[17] João Marcos Carvalho Lima and José Everardo Bessa Maia. A topical word embeddings for text classification. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2018.

[18] Alex Souza and José Everardo Bessa Maia. Agente inteligente para classificação de notícias por assunto. *Anais do Computer on the Beach*, 2019.

[19] Kássio TC Junqueira and Anita Maria da Rocha Fernandes. Análise de sentimento em redes sociais no idioma português com base em mensagens do twitter. *Anais do Computer on the Beach*, 2018.

[20] Paulo T. Guerra Raul de Araújo Lima. An analysis of the sentiment classification of short messages using word2vec. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2018.

[21] Gustavo Nogueira de Sousa, Isabelle da Silva Guimarães, Antonio Fernando Lavareda Jacob Jr, and Fábio Manoel França Lobato. Gerenciamento de Publicidades na Plataforma das Redes Sociais de Acordocomategorias de Conteúdoc. *Revista SODEBRAS*, 14(166):18–23, oct 2019. ISSN 18093957. doi: 10.29367/issn.1809-3957.14.2019.166.18. URL http://sodebras.com.br/edicoes/N166.pdf.

[22] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 1971.

[23] Douglas Cirqueira, Márcia Fontes Pinheiro, Antonio Jacob, Fábio Lobato, and Ádamo Santana. A literature review in preprocessing for sentiment analysis for brazilian portuguese social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018.

[24] William Christhie, Julio CS Reis, Fabrício Benevenuto Mirella M Moro, and Virgílio Almeida. Detecção de posicionamento em tweets sobre política no contexto brasileiro. In *7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*. SBC, 2018.