Uma Abordagem para Identificar e Monitorar *Haters* em Redes Sociais Online

Thais G. Almeida
Universidade Federal do Amazonas
Manaus, AM 69077-000
tga@icomp.ufam.edu.br

Fabíola G. Nakamura
Universidade Federal do Amazonas
Manaus, AM 69077-000
fabiola@icomp.ufam.edu.br

Eduardo F. Nakamura
Universidade Federal do Amazonas
Manaus, AM 69077-000
nakamura@icomp.ufam.edu.br

ABSTRACT

Hate speeches published and difused via online environments have the potential to cause harm and suffering to individuals, and lead to social disorder beyond cyber space. In this context, we propose a novel approach to identify and monitor groups of users which propagate such contents. As preliminary results, we detail a methodology for hate speech identification based on Information Theory quantifiers (entropy and divergence) to represent documents. The results show that our methodology overperforms techniques that use data representation, such as TF-IDF and unigrams combined to text classifiers, achieving an F1-score of 86%, 84% e 96% for classifying hate, offensive, and regular speech classes, respectively. Compared to the baselines, our proposal is a win-win solution that improves efficacy (F1-score) and efficiency (by reducing the dimension of the feature vector). The proposed solution is up to 2.27 times faster than the baseline.

KEYWORDS

Haters, Social Networks Analysis, Supervised Learning

1 CARACTERIZAÇÃO DO PROBLEMA

Nesta seção, caracterizamos o problema que é foco de nossa pesquisa quanto as suas motivações sociais e justificativas científicas.

1.1 Motivação Social

Redes Sociais Online (RSO) como Facebook, Twitter, Foursquare e Instagram tornaram-se populares por oferecerem novas formas de interação entre usuários (e.g., compartilhar conteúdo, seguir amigos, realizar *check-in*). Com a emergência do número de usuários e do conteúdo gerado por eles, RSO se tornaram uma fonte de informação para empresas, governos e pesquisadores. Cientistas sociais podem estudar em larga escala o comportamento humano relacionado a vários assuntos (e.g., doenças, economia, política). Um aspecto importante relacionado ao comportamento dos usuários são as suas opiniões, uma vez que podem ser utilizadas como base de influência para o comportamento de outros usuários [24].

A formação de comunidades de usuários com opiniões similares em RSO é um fenômeno frequente, uma vez que naturalmente indivíduos tendem a se relacionar com aqueles que compartilham interesses em comum, criando assim, uma identidade coletiva. Em razão dos seus comportamentos violentos, associação com crimes e, potencial influência para a juventude, grupos de ódio (*haters*) têm atraído atenção das autoridades e da academia [7]. Tais grupos exploram aspectos de RSO, como o anonimato e políticas frágeis de publicação de conteúdo, para disseminar mensagens de ódio (e.g., racismo, xenofobia, homofobia), recrutar novos membros, e ameaçar usuários desses meios [9].

Em particular, usuários jovens são os principais alvos de *haters*, já que representam um grupo mais facilmente afetado pelos ideais propagados [7]. Há um consenso entre as autoridades de que esses grupos devem ser analisados, a fim de monitorar atividades potencialmente prejudiciais à sociedade [8, 15]. Portanto, a identificação automática de grupos sociais de ódio se faz necessária para o eventual controle de suas atividades.

1.2 Justificativa

Por meio do estudo da dinâmica (formação e evolução) de agrupamentos de *haters*, é possível adquirir uma maior compreensão sobre a organização representada, a fim de entender quais os motivos que contribuem para a estagnação de opiniões negativas e intolerância em relação a características de indivíduos (e.g., etnia, orientação sexual, raça). Além disso, tal estudo pode revelar pontos fracos que ofereçam novas formas de combate a grupos de ódio.

Identificar *haters* é uma tarefa desafiadora, pois: (i) o grande número de usuários tornam a análise difícil de ser visualizada com recursos computacionais limitados [12]; (ii) usuários que publicam discursos de ódio tendem a disfarçar palavras ofensivas inserindo asteriscos, espaços ou substituindo caracteres por outros com sons semelhantes [19]; e (iii) o discurso de ódio por vezes é dirigido a usuários que apresentam uma interseção de "características protegidas", incluindo raça, religião e orientação sexual, por exemplo. Como cada "característica protegida" está associada a termos específicos, tal interseção em uma única vítima dificulta a identificação da agressão [6].

1.3 Definição do Problema

Em nossa pesquisa, buscamos responder se é possível identificar e monitorar automaticamente grupos de ódio em Redes Sociais Online.

1.4 Objetivos

O objetivo geral desta pesquisa é propor e demonstrar a eficácia e eficiência de uma abordagem automática para identificar e monitorar *haters* em Redes Sociais Online.

Os objetivos específicos incluem:

(1) Caracterizar *haters* do ponto de vista comportamental, a fim de compreender quais suas motivações e que fenômenos

In:XVII Workshop de Teses de Dissertações (WTD 2017), Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN 978-85-7669-380-2.

da psicologia (e.g., homofilia) colaboram para difusão de ódio em ambientes online;
(2) Propor um algoritmo eficiente e eficaz que combine técnicas de Aprendizagem de Máquina, Análise de Redes Sociais, Processamento de Sinais e Teoria da Informação, para identificar e monitorar *haters* em ambientes online;
(3) Demonstrar a eficácia e eficiência do algoritmo proposto em relação ao estado-da-arte.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, apresentamos conceitos fundamentais para a compreensão do artigo, como a definição formal de discurso de ódio, classificação de textos e quantificadores de informação.

2.1 Discurso de Ódio

Segundo Cohen-Almagor [9], o discurso de ódio é definido como malicioso, motivado por preconceito e dirigido a uma pessoa ou a um grupo de pessoas devido a algumas de suas características inatas reais ou percebidas. Tais características incluem gênero, raça, religião, etnia, origem nacional, deficiência ou orientação sexual.

2.2 Redes Sociais

Redes sociais são conjuntos de pessoas ou grupos de pessoas com algum padrão de contato ou interação entre si [22]. Matematicamente, redes sociais podem ser modeladas como grafos $G(V, E)$, onde V corresponde ao conjunto de vértices (indivíduos) e E corresponde ao conjunto de arestas (interações) do grafo [26]. Desta forma, é possível realizar análises que revelam informações implícitas como o nível de influência de vértices e a padrão de propagação de doenças em uma cidade, por exemplo.

2.3 Detecção de Comunidades

Comunidades são definidas como grupos de vértices cuja densidade das arestas que os interligam é maior do que a densidade das arestas que ligam diferentes grupos [5]. O estudo de comunidades em redes sociais nos permite compreender a organização do sistema representado, bem como sua função [3].

2.4 Classificação de Texto

O processo de inserção de documentos em classes (ou categorias), i.e., de associação de um ou mais rótulos de classe em cada documento, é chamado de classificação de textos [4]. Para esta tarefa, comumente são utilizados algoritmos de Aprendizagem de Máquina supervisionados [1, 24] e não-supervisionados [18, 19].

2.5 Quantificadores de Teoria de Informação

Quantificadores associados a estatísticas de primeira ordem de palavras e a outros elementos linguísticos têm sido empregados para quantificar o tamanho, a coerência e a distribuição de vocabulários em amostras de linguagem de vários tipos [21]. A seguir, apresentamos dois quantificadores de Teoria da Informação que foram utilizados em nossos resultados preliminares (ver subseção 5.2.2):

2.5.1 **Entropia de Shannon.** A entropia de *Shannon* pode ser definida como uma medida para quantificar a incerteza de uma distribuição p [13]. Seja x uma variável aleatória, com valores pertencentes a um conjunto finito χ, a entropia de x é formulada como:

$$S(\chi) = -\sum_{x \in \chi} p(x) \log p(x), \quad (1)$$

assim, a entropia normalizada de Shannon é dada por $H(x) = S(\chi)/S_{max}$, onde $S_{max} = \log_2 |\chi|$ é equivalente a entropia máxima.

2.5.2 **Divergência de Jensen-Shannon.** Esta divergência é definida como uma medida de distância entre duas distribuições de probabilidades [14]. Sejam p e q distribuições de probabilidades, e S a entropia de Shannon, então a divergência de Jensen-Shannon é calculada da seguinte forma:

$$JSD(p, q) = S\left(\frac{p+q}{2}\right) - \frac{S(p) + S(q)}{2}. \quad (2)$$

3 TRABALHOS RELACIONADOS

Nesta seção apresentamos uma breve revisão sobre trabalhos envolvendo identificação de discursos de ódio (*hate speech*) e de grupos sociais de ódio (*haters*) em ambientes online. Destacamos que a tarefa de identificar discursos de ódio, precede a de detectar *haters* e, portanto, faz-se necessário o estudo de técnicas para automatizar tal identificação.

3.1 Identificação de Discursos de Ódio

A identificação de discursos de ódio em ambientes online pode ser modelada como um problema de classificação de texto. Na literatura, trabalhos envolvendo classificação de textos, por meio de métodos supervisionados, utilizam duas principais formas de representação dos dados: n-gramas [18, 19] e TF-IDF [16, 24]. Na primeira abordagem, os documentos são convertidos em segmentos de sentença e são utilizados no vetor de características de classificadores. Na segunda abordagem, além de segmentar os documentos, são calculados as medidas de ponderação *Term Frequency* (TF) e *Inverse Document Frequency* (IDF) para cada palavra.

Souza et al. [24] estudaram a polaridade de *tweets* referentes ao processo de *impeachment* da então presidente do Brasil. Para isso, cada palavra dos *tweets* foi representada pelos seus valores de TF-IDF. Adicionalmente, os autores compararam a eficácia de classificadores ao inferir a polaridade de textos. Pelle and Moreira [19] investigaram discursos de ódio em textos em português provenientes do site g1.globo.com. Como contribuições, os autores apresentaram uma base de dados anotada com comentários ofensivos e não ofensivos, um sistema para rotular dados e o resultado dos classificadores *Support Vector Machine* e *Naive Bayes* ao inferir discursos de ódio, utilizando como forma de representação de dados combinações de unigramas, bigramas e trigramas.

Trabalhos mais recentes como os de Mondal et al. [15] e Postal and Nakamura [20] empregaram formas alternativas de representação de dados. Mondal et al. [15] utilizam o formato de sentença "`I <intensity> <userintent> <hatetarget>`" para identificar discursos de ódio. Em tal formato, I é referente ao usuário que escreveu o comentário, `<intensity>` revela o sentimento do usuário (geralmente um verbo), `<userintent>` está relacionado a palavras que remetem ódio, `<hatetarget>` corresponde a quem o ódio está sendo dirigido. Já Postal and Nakamura [20] utilizaram distribuições de frequências de palavras para calcular a entropia de Shannon

e divergência de Jensen-Shannon [21], a fim caracterizar e classificar textos de *chats* de pedofilia. Como classificador, os autores utilizaram o *Support Vector Machine*, obtendo valor para a medida F1 de aproximadamente 90%.

3.2 Identificação de Grupos de Ódio

Um dos trabalhos pioneiros no estudo sobre comunidades de *haters* foi o de Chau and Xu [7], no qual os autores propuseram uma abordagem semi-automática que combinou técnicas de *blog spidering* e análise de redes sociais para monitorar comunidades de *haters*. Foram realizadas três tipos de análises: topológica, de centralidade e de comunidades em cerca de 28 blogs, totalizando 820 usuários. Dentre seus resultados, os autores destacaram que: a rede composta pelos usuários é descentralizada (número isolado de *clusters*); alguns usuários atuam como líderes de opiniões ou pontes de comunicação na rede; e que *haters* representam um problema internacional, já que não se concentram em um país ou região específica.

Chen et al. [8] e Sureka et al. [25] estudaram a presença de *haters* em sites de multimídias. Sureka et al. [25] utilizaram técnicas de mineração de dados e análise de redes sociais para identificar vídeos, usuários e comunidades de ódio no site Youtube. Como resultado, os autores destacaram que por meio de sua abordagem é possível descobrir usuários centrais e influentes, bem como comunidades ocultas de *haters*. Já Chen et al. [8] conduziram uma análise exploratória para verificar a presença de grupos Jihadistas em ambientes online (Youtube, Second Life, Blogs). Os autores argumentaram que sites com uma larga base de dados, popularidade entre os jovens e que suportem vários tipos de mídias são frequentemente alvos de grupos extremistas. Como resultado, os autores destacam que 28 blogs, 80 vídeos do Youtube e 7 grupos (23 a 228 integrantes) do Second Life apresentaram conteúdo extremista.

Em O'Callaghan et al. [17], os autores investigaram o potencial de diferentes redes sociais (Twitter, Facebook, Youtube) como ponte entre comunidades de ódio que propagam conteúdo de extrema direita. Para tanto, é apresentada uma abordagem que combina dados heterogêneos de diferentes redes sociais, para construir uma rede homogênea. Como resultado, os autores destacaram a importância de diferentes bases de dados na identificação de comunidades, que não teriam sido identificadas se apenas uma rede social tivesse sido considerada. Já em Ferrara et al. [11], os autores apresentaram uma abordagem que combina aprendizagem de máquina, redes e atributos temporais para (i) detectar usuários extremistas, (ii) prever se usuários regulares se tornarão adeptos de ideologias extremistas, e (iii) prever se usuários regulares manterão contato com usuários extremistas. Para tanto, foram utilizados os métodos *Random Forest* e *Logistic Regression*. Como resultado, os autores destacaram que sua abordagem se mostrou promissora por apresentar valores de AUC de 93%, 80% e 72% nas tarefas anteriores.

4 CONTRIBUIÇÕES ESPERADAS

O foco desta pesquisa consiste em propor e demonstrar a eficácia de uma abordagem automática para identificar e monitorar grupos de ódio em Redes Sociais Online. Para isso, pretendemos utilizar os trabalhos de Postal and Nakamura [20] e Chen et al. [8] como *baselines* para as etapas de identificação de discursos e grupos de ódio, respectivamente. Em relação ao trabalho de Postal and Nakamura [20], utilizaremos quantificadores de Teoria da Informação para caracterizar comentários online e técnicas de Aprendizagem de Máquina para identificar se tais comentários contêm discursos de ódio ou não. Quanto ao trabalho de Chen et al. [8], pretendemos utilizar conceitos de Redes Complexas para identificar comunidades de usuários que propagam conteúdos de ódio.

Como contribuições esperadas da nossa pesquisa para a academia, indústria e sociedade, pretendemos:

- Apresentar um *survey* sobre os trabalhos correlatos.
- Fornecer um algoritmo eficiente e eficaz para identificar e monitorar discursos e grupos de ódio em ambientes online;
- Caracterizar usuários *haters* do ponto de vista comportamental, a fim de compreender quais suas motivações e que fenômenos da psicologia colaboram para difusão de ódio em ambientes online;
- Elaborar estratégias eficazes de combate a grupos sociais de ódio, por meio da análise topológica das suas comunidades.

Destacamos que o nosso trabalho apresentará como diferenciais: (i) a utilização de distribuições de TF-IDF de palavras, para calcular quantificadores de Teoria da Informação na etapa de identificação de discurso de ódio e (ii) o uso de grafos sociais em conjunto com técnicas de processamento de sinais, a fim de identificar e monitorar grupos de ódio. Quanto ao item (i), adaptaremos a abordagem proposta em Postal and Nakamura [20], onde os autores utilizam somente distribuições de frequências de palavras para representar informações. Segundo Shuman et al. [23], dados presentes em grafos consistem em uma coleção finita de amostras que podem ser interpretadas como um sinal discreto no domínio de tempo. Desta forma, em relação ao item (ii), pretendemos monitorar a dinâmica dos grupos de ódio por meio de séries temporais, nas quais os sinais seriam equivalentes as opiniões desses usuários.

5 ESTADO ATUAL DO TRABALHO

Atualmente, nosso trabalho está concentrado nas tarefas de coleta de dados e identificação de discursos de ódio. As subseções seguintes descrevem cada uma dessas tarefas.

5.1 Coleta de Dados

Estamos construindo uma base de dados que contém comentários de usuários do site de notícias g1.globo.com. Escolhemos este site devido à sua popularidade, integração com Redes Sociais Online (e.g., Twitter, Facebook), e estruturação dos comentários, uma vez que é possível capturar subconversas de usuários (ver figura 1). Assim, pretendemos modelar, por meio de grafos sociais, a interação entre usuários (*haters* e regulares) a partir da estrutura de seus comentários.

Até o momento, coletamos 56.096 comentários escritos em português sobre diferentes tópicos (e.g., política, esportes, economia). Com a ajuda de voluntários, pretendemos rotular manualmente uma fração expressiva desses dados como apresentando discursos de ódio na forma de: **xenofobia**, **intolerância religiosa**, **homofobia**, **sexismo**, **ideologia política**, **rivalidade** e **inveja**. Ressaltamos que diferentes formas de discurso de ódio podem estar presentes em um mesmo comentário, o que torna mais desafiadora a tarefa de classificação de textos.

Figura 1: Exemplo da estrutura de conversas do site *g1.globo.com*. O encadeamento de usuários indica uma subconversa entre eles.

5.2 Identificação de Discursos de Ódio

Para a tarefa de identificação de discurso de ódio, propusemos uma abordagem composta de três etapas principais (ilustradas na figura 2): (i) pré-processamento dos dados, que consiste na remoção de ruídos; (ii) extração de características, onde são computados os quantificadores de Teoria da Informação; e (iii) classificação, que consiste na inferência de discursos de ódio.

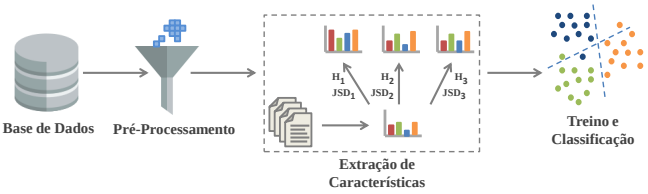

Figura 2: Etapas da abordagem proposta.

5.2.1 **Pré-processamento.** Nesta etapa, os textos pertencentes a base de dados foram submetidos a uma filtragem, a fim de remover conteúdos com pouco valor semântico. Desta forma, removemos URLs, pontuações, dígitos, menções e *stopwords*. Logo em seguida, submetemos os textos a um processo de segmentação em palavras (tokenização). Assim, cada texto da base de dados foi convertido em uma lista de segmentos.

5.2.2 **Extração de Características.** Nesta etapa, primeiramente computamos os valores de TF-IDF das palavras do vocabulário. Em seguida, calculamos dois tipos de histogramas: referência e individual. O histograma de referência considera os valores de TF-IDF das palavas do vocabulário por tipo discurso, isto é, para cada rótulo do problema de classificação de texto, será computado um histograma de referência. O histograma individual é calculado para cada documento da base de dados, isto é, considera somente os valores de TF-IDF das palavras que compõe o documento individualmente. Após a definição dos histogramas, utilizamos os quantificadores entropia e divergência (descritos na subseção 2.5) para calcular os atributos que representam cada instância da base de dados no vetor de características de classificadores.

No contexto de classificação de textos, a entropia normalizada de Shannon quantifica a riqueza do vocabulário utilizado, isto é, valores de entropia próximos a 0 indicam que as palavras do vocabulário possuem frequências baixas, enquanto valores próximos a 1 indicam uma distribuição mais uniforme dessas frequências. Por outro lado, a divergência de Jensen-Shannon indica quão similar um texto é em relação a um determinado tipo de discurso. Assim, valores de divergência próximos a 0 indicam similaridade e próximos a 1 indicam dissimilaridade.

Como resultado final do processo de extração de características, teremos dois atributos: uma entropia e uma divergência. Cada um desses dois atributos é calculado para cada classe do problema, isto é, se o problema possuir três classes haverão seis atributos representando cada documento da base de dados.

Referenciaremos nossa abordagem para identificação de discursos de ódio como **H+JSD com TF-IDF**.

5.2.3 **Treino e Classificação.** Nesta etapa, utilizamos métodos supervisionados de aprendizagem de máquina para classificar os textos da base de dados como: discurso de ódio, ofensivo e regular.

6 METODOLOGIA EXPERIMENTAL

Esta seção descreve a metodologia experimental utilizada na tarefa de identificação de discursos de ódio.

6.1 Base de Dados

Como a base de dados descrita na seção 5.1 está em processo de construção, utilizamos a base rotulada proposta por Davidson et al. [10] e publicamente disponível[1]. Tal base é composta por 14.442 textos curtos escritos em inglês provenientes da rede social Twitter. Em sua elaboração, cada texto foi rotulado por três ou mais especialistas dentre as classes: discursos de ódio, ofensivo e regular (ver exemplos na tabela 1). Para decidir o rótulo de um *tweet*, foi utilizado o critério de voto majoritário. Em caso de empate, o *tweet* era descartado. A taxa de concordância dos especialistas foi de 92%.

Tabela 1: Exemplos de *tweets* para cada classe.

Exemplo de *tweet*	Classe
`These faggot ass niggas really think they a female`	ódio
`I'm never gonna be ok with my nigga around alot of bitches while with his boys. Cause I was once that female your boys put you on !!`	ofensivo
`There may be radical Muslims but there are radical Christian's, Jews, other religions and indeed atheists. #TheLostMuslim`	regular

É importante ressaltar que discurso de ódio foi definido como uma linguagem que é usada para expressar ódio direcionado a grupos de indivíduos, isto é, para depreciar, humilhar ou insultar seus membros. Por outro lado, o discurso ofensivo foi caracterizado

[1] https://www.crowdflower.com/wp-content/uploads/2016/03/twitter-hate-speech-classifier-DFE-a845520.csv

Tabela 2: Comparação da eficácia dos métodos K-Nearest Neighbor, Max Entropy e Naive Bayes utilizando diferentes formas de representação. As medidas de avaliação são precisão, revocação e F1.

		K-Nearest Neighbor (KNN)			Max Entropy (ME)			Naive Bayes (NB)		
Características	Classe	Precisão	Revocação	F1	Precisão	Revocação	F1	Precisão	Revocação	F1
H+JSD (com TF-IDF)	Ódio	**0,84±0,001**	**0,88±0,002**	**0,86±0,002**	0,48±0,003	0,95±0,023	0,63±0,008	0,52±0,001	0,90±0,022	0,66±0,001
	Ofensivo	**0,86±0,004**	**0,82±0,013**	**0,84±0,004**	0,45±0,009	0,44±0,115	0,44±0,043	0,42±0,022	0,52±0,083	0,46±0,043
	Regular	**0,96±0,018**	**0,97±0,022**	**0,96±0,019**	0,95±0,084	0,03±0,136	0,06±0,086	0,94±0,083	0,01±0,131	0,03±0,095
H+JSD (somente TF)	Ódio	**0,78±0,002**	**0,84±0,001**	**0,81±0,001**	0,62±0,000	0,85±0,000	0,72±0,000	0,56±0,001	0,67±0,001	0,61±0,001
	Ofensivo	**0,78±0,001**	**0,76±0,019**	**0,77±0,010**	0,85±0,050	0,59±0,059	0,69±0,005	0,49±0,014	0,65±0,005	0,56±0,011
	Regular	**0,87±0,014**	**0,83±0,013**	**0,85±0,011**	0,83±0,036	0,78±0,040	0,80±0,016	0,70±0,031	0,34±0,055	0,46±0,023
TF-IDF	Ódio	0,50±0,012	0,68±0,021	0,57±0,013	**0,62±0,011**	**0,55±0,028**	**0,58±0,018**	0,58±0,011	0,61±0,025	0,60±0,016
	Ofensivo	0,48±0,024	0,49±0,025	0,48±0,022	**0,51±0,015**	**0,61±0,025**	**0,55±0,016**	0,51±0,018	0,56±0,024	0,54±0,017
	Regular	0,83±0,016	0,51±0,017	0,63±0,016	**0,86±0,017**	**0,79±0,024**	**0,83±0,019**	0,83±0,021	0,70±0,027	0,76±0,019
Unigrama	Ódio	0,55±0,020	0,72±0,030	0,62±0,018	**0,61±0,021**	**0,65±0,019**	**0,63±0,016**	0,58±0,020	0,63±0,025	0,61±0,021
	Ofensivo	0,52±0,031	0,41±0,030	0,45±0,021	**0,52±0,018**	**0,54±0,024**	**0,53±0,019**	0,51±0,017	0,56±0,018	0,54±0,016
	Regular	0,80±0,016	0,70±0,028	0,75±0,015	**0,85±0,021**	**0,76±0,027**	**0,80±0,021**	0,87±0,014	0,72±0,018	0,79±0,015

como aquele que pode conter palavras depreciativas em relação a um grupo, porém é utilizado de uma maneira qualitativa diferente, isto é, o contexto é considerado.

6.2 Classificadores de Texto

Neste trabalho utilizamos os seguintes classificadores de texto: *Multinomial Naive Bayes*, *K-Nearest Neighbor* (K=5) e *Maximum Entropy*. Para o classificador *K-Nearest Neighbor*, escolhemos empiricamente o melhor número de vizinhos considerados para a inferência. As implementações dos métodos supervisionados considerados neste trabalho estão disponíveis na API *Scikit learn*[2].

Para garantir a validação estatística dos resultados gerados por meio da classificação, treinamos os métodos utilizando a técnica de validação cruzada 10-*ten fold*. Esta técnica consiste em construir, 10 vezes, dez diferentes classificadores c_1,c_2...,c_{10} e dividir o conjunto de documentos de treinamento em dez partições distintas (*folds*) de tamanho: N_1, N_2,..., N_{10}. Desta forma, os classificadores utilizam o i-ésimo *fold* como conjunto de teste, e os documentos restantes como o conjunto de treinamento [4]. Utilizamos a validação cruzada 10-*ten fold* estratificada que respeita a proporção das classes em cada *fold*. Os resultados coletados são complementados com intervalos de confiança para o nível $\alpha = 95\%$.

6.3 Métricas de Avaliação

Para avaliar a eficácia dos classificadores na tarefa de identificação de discursos de ódio, utilizamos as seguintes métricas: precisão, revocação e F1. Enquanto a precisão consiste na fração de documentos atribuídos a uma determinada classe que realmente pertencem a esta classe (segundo o conjunto de teste), a revocação representa a fração de todos os documentos que pertencem a uma determinada classe (segundo o conjunto de teste) e foram atribuídas corretamente a esta classe pelo classificador. Já a métrica F1 pode ser definida como uma medida que busca relacionar as métricas de precisão e revocação a fim de obter uma medida de qualidade que equilibre a importância relativa destas duas métricas [4].

[2]http://scikit-learn.org/stable/

7 RESULTADOS PRELIMINARES

Nesta seção apresentamos uma análise quantitativa entre a abordagem de representação de dados proposta neste trabalho e as formas de representação apresentadas na seção 3.1: unigrama, TF-IDF e H+JSD [20]. Para isso, combinamos os três classificadores escolhidos com cada uma das representações, e os avaliamos em relação as métricas descritas na subseção 6.3. A tabela 2 apresenta os resultados da nossa avaliação. Destacamos que parte de nossos resultados preliminares, estão descritos em Almeida et al. [2].

O classificador que obteve maior eficácia utilizando a abordagem proposta neste trabalho (H+JSD com TF-IDF) foi o KNN, que alcançou valor para a medida F1 de $0,86 \pm 0,002$, $0,84 \pm 0,004$ e $0,96 \pm 0,019$ para as classes ódio, ofensivo e regular, respectivamente. Na tabela 2 é possível observar que o classificador KNN apresentou maior dificuldade em discriminar as classes referentes a discursos de ódio e ofensivo. Isto é explicado pela interseção de palavras pejorativas que ambas as classes apresentam.

Em relação a forma de representação H+JSD, o classificador KNN novamente apresentou os melhores valores para medida F1: $0,81 \pm 0,001$, $0,77 \pm 0,010$ e $0,85 \pm 0,011$ para as classes ódio, ofensivo e regular, respectivamente. Nesta abordagem, o classificador *Naive Bayes* (NB) apresentou uma melhora nos seus valores para medida F1, com exceção da classe regular que obteve valor de revocação $0,34 \pm 0,055$ o que contribuiu para seu baixo valor de F1.

Para a representação TF-IDF, o classificador que obteve melhor eficácia foi o *Maximum Entropy* (ME), que alcançou valores de F1 de $0,58 \pm 0,018$, $0,55 \pm 0,016$ e $0,83 \pm 0,019$ para as classes ódio, ofensivo e regular, respectivamente. É importante observar que nesta abordagem, nenhum dos classificadores conseguiu discriminar de forma eficaz as classes ódio e ofensiva.

Para a representação utilizando unigramas, o classificador ME alcançou os melhores valores para a medida F1 em relação as classes ódio e regular, enquanto o classificador NB apresentou melhor valor para a classe ofensiva. Novamente, observamos que a maior dificuldade dos classificadores está em discriminar a classe ofensiva.

A melhor solução foi a solução proposta, H+JSD com TF-IDF, utilizando o KNN. Este resultado sugere que a inclusão de ambos

os pesos TF e IDF na construção dos histogramas particulares e de referência, auxiliou em uma maior capacidade de generalização em relação ao H+JSD proposto por Postal and Nakamura [20].

Em relação a eficiência, na abordagem proposta, **H+JSD com TF-IDF**, cada documento da base de dados é representado por um conjunto de apenas seis atributos (uma entropia e uma divergência para cada classe), contra mais de 20.100 (tamanho do vocabulário) para o TF-IDF tradicional. Como consequência, considerando o tempo de classificação, o H+JSD com TF-IDF apresentou em nossos experimentos um *Speedup* de 2,27 em relação ao TF-IDF tradicional (considerando o KNN).

8 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma proposta de abordagem para identificar e monitorar *haters*. Como resultados parciais, apresentamos a coleta de uma base de dados com 56.096 comentários e uma abordagem que utiliza quantificadores (entropia e divergência) de Teoria da Informação para representação de comentários, e posterior, identificação de discursos de ódio presentes neles. Como diferencial, destacamos a utilização dos pesos TF-IDF de palavras no cálculo de tais quantificadores.

Em comparação com o *bag of words* tradicional, baseado em TF-IDF, nossa abordagem, H+JSD com TF-IDF, chegou a apresentar um ganho superior a 15% em relação à métrica F1 (qualidade da classificação). Além disso, considerando o tempo de execução, a abordagem proposta chegou a ser 2,27 vezes mais veloz que o referido *baseline*. O ganho de eficiência é resultado da redução de dimensões que, na solução proposta, são sempre apenas seis características.

Como trabalhos futuros, pretendemos avaliar a utilização de outros quantificadores de informação na representação de textos (e.g., distância de Jaccard, coeficiente de Sørensen). Além disso, pretendemos validar nossa abordagem de identificação de discursos de ódio em outras bases de dados, tais como as apresentadas por Pelle and Moreira [19]. Quanto a base de comentários provenientes do site g1.globo.com que estamos construindo, pretendemos rotular uma fração de seus comentários e mapear as interações entre usuários em grafos sociais, a fim de monitorar e compreender de que forma se dá a dinâmica de grupos de ódio em ambientes online.

REFERÊNCIAS

[1] Thais G. Almeida, Bruno A. Souza, Alice A. F. Menezes, Carlos Figueiredo, and Eduardo F. Nakamura. 2016. Sentiment Analysis of Portuguese Comments from Foursquare. In *Proc. of the 22nd Brazilian Symposium on Multimedia and the Web*. 355–358.

[2] Thais G. Almeida, Bruno A. Souza, Fabíola G. Nakamura, and Eduardo F. Nakamura. 2017. Detecting Hate, Offensive, and Regular Speech in Short Comments. In *Proc. of the 23rd Brazilian Symposium on Multimedia and the Web*.

[3] Babak Amiri, Liaquat Hossain, John Crawford, and Rolf Wigand. 2013. Community Detection in Complex Networks: Multi–objective Enhanced Firefly Algorithm. *Knowledge-Based Systems* (2013), 1–11.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2013. *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora. 328–329 pages.

[5] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D Hwang. 2006. Complex Networks: Structure and Dynamics. *Physics reports* (2006), 175–308.

[6] Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5 (2016), 1–15.

[7] Michael Chau and Jennifer Xu. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *Int'l Journal of Human-Computer Studies* 65 (2007), 57–70.

[8] Hsinchun Chen, Sven Thoms, and Tianjun Fu. 2008. Cyber extremism in Web 2.0: An exploratory study of international Jihadist groups. In *Proc. of the Int'l Conf. on Intelligence and Security Informatics*. 98–103.

[9] Raphael Cohen-Almagor. 2011. Fighting hate and bigotry on the Internet. *Policy & Internet* 3 (2011), 1–26.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proc. of the 11th Int'l AAAI Conf. on Web and Social Media*. 512–515.

[11] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In *Int'l Conf. on Social Informatics*. 22–39.

[12] Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and Athanasios V. Vasilakos. 2013. Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine* 51 (2013), 144–150.

[13] Annick Lesne. 2014. Shannon Entropy: A Rigorous Notion at The Crossroads Between Probability, Information Theory, Dynamical Systems and Statistical Physics. *Mathematical Structures in Computer Science* 24 (2014), 240–311.

[14] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*. 880–889.

[15] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proc. of the 28th Conf. on Hypertext and Social Media*. 85–94.

[16] Felipe Moraes, Marisa Vasconcelos, Patrick Prado, Jussara Almeida, and Marcos Gonçalves. 2013. Polarity analysis of micro reviews in foursquare. In *Proc. of the 19th Brazilian Symposium on Multimedia and the Web*. 113–120.

[17] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2013. Uncovering the wider structure of extreme right communities spanning popular online networks. In *Proc. of the 5th Annual ACM Web Science Conf.* 276–285.

[18] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of the ACL-02 Conf. on Empirical Methods in Natural Language Processing*. 79–86.

[19] Rogers P. Pelle and Viviane P. Moreira. 2017. Offensive Comments in the Brazilian Web: a dataset and baselines results. In *Proc. of the 6th Brazilian Workshop on Social Network Analysis and Mining*. 1–160.

[20] Juliana G. Postal and Eduardo F. Nakamura. 2017. Utilizando Teoria da Informação para Identificar Conversas de Pedofilia em Redes Sociais de Mensagens Instantâneas. In *Proc. of the 14th Simpósio Brasileiro de Sistemas Colaborativos*. 1–345.

[21] Osvaldo A. Rosso, Hugh Craig, and Pablo Moscato. 2009. Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications* 388 (2009), 916–926.

[22] John Scott. 2012. *Social network analysis*. Sage.

[23] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* 30 (2013), 83–98.

[24] Bruno A. Souza, Thais G. Almeida, Alice A. F. Menezes, Fabíola G Nakamura, Carlos Figueiredo, and Eduardo F. Nakamura. 2016. For or Against?: Polarity Analysis in Tweets about Impeachment Process of Brazil President. In *Proc. of the 22nd Brazilian Symposium on Multimedia and the Web*. 335–338.

[25] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. 2010. Mining youtube to discover extremist videos, users and hidden communities. *Information retrieval technology* (2010), 13–24.

[26] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge University Press.