Anais do XXV Encontro de Iniciação Científica – ISSN 1982-0178
Anais do X Encontro de Iniciação em Desenvolvimento Tecnológico e Inovação – ISSN 2237-0420
22 a 26 de setembro de 2020

IDENTIFICAÇÃO DE LINGUAGEM OFENSIVA EM MÍDIAS SOCIAIS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Murilo de Paula Araújo
Faculdade de Engenharia de Computação
CEATEC
murilo.pa@puccampinas.edu.br

Juan Manuel Adán Coello
Grupo de Pesquisa em Sistemas Inteligentes
(GPqSI) CEATEC
juan@puc-campinas.edu.br

Resumo: *À medida que cresce o número de pessoas que interagem através das mídias sociais, multiplicam-se os casos de agressão e abuso online. No trabalho descrito neste artigo, o problema da detecção de linguagem ofensivo em postagens publicadas nessas mídias foi tratado utilizando comitês de classificadores baseados em redes neurais convolucionais. Nos experimentos realizados para avaliar esta abordagem, constatou-se que ela permitiu alcançar desempenho superior a outra proposta apresentada na literatura recente, para os mesmos conjuntos de dados.*

Palavras-chave: *Discurso de ódio, redes neurais profundas.*

Área do Conhecimento: *Ciências Exatas e da Terra – Ciência da computação.*

1. INTRODUÇÃO

À medida que cresce o número de pessoas que interagem através das mídias sociais tem sido relatado um aumento expressivo de casos de agressão *online*, incluindo assédios de diversas naturezas e a presença de discurso de ódio. O alcance e a extensão da Internet deram a esses incidentes um poder e uma influência sem precedentes, afetando a vida de bilhões de pessoas. Casos de agressão e abuso *online*, como *cyberbullying ou bullying digital*, tem criado problemas de várias gravidades aos usuários, levando muitos deles à desativação de contas, autoflagelação e, em casos extremos, até mesmo ao suicídio. Portanto, incidentes de comportamento agressivo *online* tornaram-se uma importante fonte de conflito social, com potencial de produzir atividades criminosas [1].

Identificar linguagem ofensiva em postagens em mídias sociais manualmente é uma tarefa impraticável em função das limitações físicas e cognitivas dos seres humanos para lidar com grandes quantidades de informação. Em adição, trata-se de tarefa difícil de executar com imparcialidade, visto que as pessoas frequentemente prestam mais atenção a opiniões que são consistentes com suas próprias ideias, crenças e preferências [2]. Em adição, em muitos casos, o que é considerado ofensivo por uma pessoa pode não ser para outra.

O problema de identificar linguagem ofensiva em textos pode ser abordado utilizando métodos de classificação de texto, empregados para tratar problemas semelhantes em trabalhos anteriores do Grupo de Pesquisa em Sistemas Inteligentes da PUC-Campinas, tais como a análise de sentimento [3] e a detecção de posicionamento [4].

Embora semelhantes, a análise de sentimento e a detecção de posicionamento têm características específicas. A análise de sentimento visa determinar se um texto expressa um sentimento positivo, negativo ou neutro [5]. Na detecção de posicionamento, deve-se determinar se o texto é favorável, desfavorável ou neutro com relação a um objeto de interesse pré-escolhido. Este objeto (ou alvo) pode inclusive não ser mencionado no texto e nem ser o alvo do texto [6].

A detecção de linguagem ofensiva envolve aspectos das duas tarefas precedentes, mas não se confunde com elas. Um texto pode expressar uma posição não favorável com relação a algum grupo ou pessoa, ou um sentimento negativo, sem, no entanto, ser ofensivo ou incitar ao ódio contra essa pessoa ou grupo.

Diversas são as abordagens usados na área de Processamento de Língua Natural para tratar de problemas de classificação de texto. Entre elas, destacam-se na atualidade os métodos de aprendizado de máquina, em especial as Redes Neurais Convolucionais (RNC), *Convolutional Neural Networks* em inglês, empregadas em trabalhos anteriores do Grupo de Pesquisa em Sistemas Inteligentes da PUC-Campinas [3],[4],[7]. Em um desses trabalhos foram criados comitês (*ensembles*) compostos por redes neurais convolucionais para a determinação do posicionamento expresso em postagens de mídias sociais [7]. Essa mesma abordagem foi utilizada no trabalho descrito neste artigo para a detecção de linguagem ofensiva.

2. REDES NEURAIS

Uma rede neural pode ser entendida com um conjunto de unidades computacionais (também denominadas de neurônios artificiais, células ou nós) e um conjunto de conexões direcionadas entre elas. Cada unidade tem a capacidade de realizar a leitura de suas entradas, através das conexões, e computar um valor para a sua saída, chamado de valor de ativação, que pode ser a entrada de outros neurônios na sequência [8]. As conexões podem possuir um peso associado, definido durante a etapa de treinamento ou aprendizagem da rede, através de um processo de minimização do erro das unidades da camada final, denominada de camada de saída.

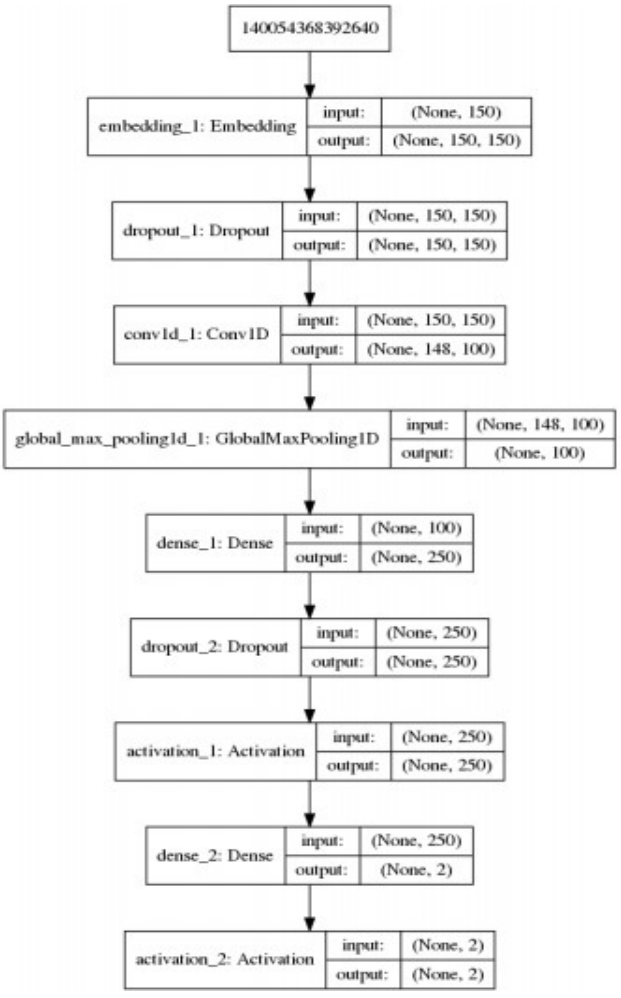

Figura 1: Rede Neural Convolucional utilizada.

As redes neurais convolucionais são redes neurais que usam convoluções em sua estrutura. Redes com esta arquitetura inicialmente se destacaram na área de processamento de imagens (classificação e reconhecimento), entretanto ultimamente elas vêm sendo exploradas também no processamento de língua natural, com resultados ao nível do estado da arte [9][10].

Uma camada de convolução geralmente é composta por vários filtros (funções janela) visando extrair características dos dados sobre os quais são aplicados.

Para poder utilizar as redes neurais, é necessário realizar antes um processo de treinamento. Durante esse processo, são utilizados os dados disponíveis (exemplos) para realizar o ajuste de seus parâmetros, inclusive dos filtros no caso das Redes Neurais Convolucionais.

Nos experimentos foi utiliza a RNC com a arquitetura mostrada na Figura 1, empregada anteriormente em [3] e [7].

3. COMITÊS DE CLASSIFICADORES

Um *ensemble,* ou comitê de classificadores, consiste de um grupo de classificadores cujas saídas são combinadas com o intuito de obter uma classificação única, tendo como principal objetivo o aumento da acurácia, relativamente aos classificadores individuais que o compõe. Há vários métodos para a construção de *ensembles* [11] [12]. Neste trabalho, foi usado o método de votação, onde é selecionada a classificação comum ao maior número de classificadores.

4. *DATASETS*

Para a avaliação do desempenho da abordagem adotada e comparação com outras abordagens foram utilizados três conjuntos de dados (*datasets*) pré-anotados, dois e, língua portuguesa e um em língua inglesa, de diferentes fontes, contextos e com características variadas, a saber: OffComBr-2 e OffComBr-3 e kaggle-test. O conjunto de dados *OffComBr2*[1] consiste de comentários coletados de *g1.globo.com,* o *site* de notícias mais acessado do Brasil. Contém 1250 comentários (instâncias) escritos em português, sendo 419 deles classificados como ofensivos (33,52% do total) e 831 classificados como não ofensivos (66,48% do total). Cada comentário foi anotado por três juízes, que indicaram se o comentário era ofensivo ou não. No caso de uma resposta afirmativa, o anotador tam-

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

bém categorizou a ofensa como racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou xingamento. A classe associada a cada comentário (ofensivo ou não ofensivo) foi a escolhida por ao menos dois juízes.

O conjunto *OffComBr-3* foi coletado da mesma fonte, mas é mais restrito que o *OffComBr-2* por conter apenas comentários em que houve concordância entre os três juízes. O conjunto contém 1033 comentários, sendo 201 classificados como ofensivos (19,46% do total) e 832 classificados como não ofensivos (80,54% do total).

O conjunto *kaggle*[2] consiste de comentários extraídos de conversas *online*, cada um classificado para refletir se pode ou não ser considerado ofensivo a outro participante da conversa. Assim, cada comentário é rotulado como ofensivo ou não ofensivo. O conjunto foi utilizado na competição *Kaggle* de 2012 e está dividido em dois subconjuntos, o *kaggle-train,* com 3947 instâncias, cada uma consistindo do texto do comentário e o rótulo associado, e o subconjunto *kaggle-test,* que foi usado nos experimentos descritos neste texto, contendo 2647 instâncias, sendo 693 instâncias classificadas como ofensivas (26,18% do total) e 1954 instâncias classificadas como não ofensivas (73,82% do total).

5. AVALIAÇÃO EXPERIMENTAL

Na avaliação descrita foi utilizada o método de validação cruzada em cinco etapas, de forma que todos os dados disponíveis são utilizados para treinamento e também para validação, sendo 4/5 do *dataset* utilizado para o treinamento dos classificadores e 1/5 para a sua validação, em processo repetido cinco vezes. Os valores das métricas usadas para avaliar o desempenho dos classificadores apresentados representam as médias das cinco execuções.

Os experimentos foram executados em um *notebook* com processador Intel i3-6100U 2.3.Ghz, 4 GB de memória principal, utilizando o UBUNTU, com a máquina em *dual boot*.

Inicialmente, foram realizados alguns experimentos com o *kaggle-test*, a fim de determinar o número de classificadores a incluir no comitê que propiciasse uma boa relação entre aumento de acurácia e o tempo de execução.

Foram criados comitês contendo 5, 7, 25, 49, 98 e 127. A Tabela 1 mostra a acurácia média dos classificadores componentes desses comitês.

Tabela 1: Acurácia média dos classificadores componentes dos comitês, em função do número de classificadores que os compõem - *kaggle-test dataset*

Classificadores	Acurácia Média (%)
5	90,464
7	90,706
25	90,563
49	90,490
98	90,354
127	90,525

De forma análoga, a Tabela 2 mostra a acurácia dos comitês em função do número de classificadores que os compõem.

Tabela 2: Acurácia dos comitês em função do número de classificadores que os compõem - *kaggle-test dataset*

Classificadores	Acurácia (%)
5	91.128
7	91.209
25	91.484
49	90.938
98	91.357
127	91.467

Nos experimentos foi também verificado que o tempo de execução do comitê aumenta de maneira linear, de modo que a cada uma nova instância da rede neural convolucional incluída no comitê o seu tempo de execução aumenta em aproximadamente um minuto. O aumento do tempo de execução é também acompanhado de maior consumo de memória, que pode levar a exaurir os recursos disponíveis e levar ao travamento da execução, o que ocorreu, por exemplo, quando se tentou executar um comitê com 194 classificadores.

[2] https://www.kaggle.com/c/detecting-insults-in-social-commentary/data

A partir dos resultados descritos, foram realizados novos experimentos, utilizando agora outros dois conjuntos de dados: OffComBr2 e OffComBr3, em adição ao *kaggle-test*. Foi empregado um comitê com sete classificadores, por ter apresentado uma boa relação entre acurácia e tempo de execução nos experimentos anteriores.

Para uma melhor visualização, os resultados destes experimentos serão apresentados sob a forma de matrizes de confusão, a partir das quais serão calculadas as métricas acurácias, *Medida F* e *ROC-AUC*.

As matrizes de confusão resultantes deste novo conjunto de experimentos são apresentadas nas Tabelas 3, 4 e 5, e os valores das métricas calculadas a partir dos resultados obtidos são apresentados na Tabela 6

Tabela 3: Matriz de confusão – comitê com 7 classificadores e conjunto de dados *OffComBr2*

Matriz de Confusão			
		Resposta Esperada	
		Ofensivo	Não Ofensivo
Resposta Prevista	Ofensivo	410 Verdadeiro Positivo	14 Falso Positivo
	Não Ofensivo	7 Falso Negativo	817 Verdadeiro Negativo

Tabela 4: Matriz de confusão – comitê com 7 classificadores e conjunto de dados *OffComBr3*

Matriz de Confusão			
		Resposta Esperada	
		Ofensivo	Não Ofensivo
Resposta Prevista	Ofensivo	201 Verdadeiro Positivo	14 Falso Positivo
	Não Ofensivo	0 Falso Negativo	818 Verdadeiro Negativo

Tabela 5: Matriz de confusão – comitê com 7 classificadores e conjunto de dados *kaggle-test*

Matriz de Confusão			
		Resposta Esperada	
		Ofensivo	Não Ofensivo
Resposta Prevista	Ofensivo	689 Verdadeiro Positivo	19 Falso Positivo
	Não Ofensivo	4 Falso Negativo	1935 Verdadeiro Negativo

Os resultados destes experimentos foram comparados com os obtidos em [13], que propõe uma abordagem denominada Hate2Vec, que consiste de um comitê composto pelos seguintes classificadores de base: (i) um classificador baseado em léxico, com o uso de vetorização de palavras (*word embedding);* (ii) um classificador do tipo regressão logística, com o uso de vetorização de comentários (*comment embeddings*) e (iii) um classificador SVM em que comentários são representados usando *bag-of-words (BOW).* Os resultados da aplicação dessa abordagem são apresentados na Tabela 7.

Tabela 6: Resultados obtidos utilizando comitê composto de sete classificadores baseados em RNCs

	Medida F	ROC-AUC	Acurácia
OffComBr2	0,98	0,97	0,98
OffComBr3	0,97	0,96	0,98
Kaggle-test	0,93	0,89	0,91

Tabela 7: Resultados obtidos utilizando Hate2Vec

	Medida F	ROC-AUC	Acurácia
OffComBr2	0,97	0,98	0,97
OffComBr3	0,94	0,94	0,94
Kaggle-test	0,91	0,88	0,91

A comparação dos resultados apresentados nas Tabelas 7 e 8 permite constatar que o comitê de classificadores foco deste trabalho, apresenta desempenho superior ao alcançado pela abordagem Hate2Vec em todos os cenários avaliados, exceto um em que apresenta desempenho igual (acurácia para o conjunto *kaggle-test*).

6. CONCLUSÃO

Os experimentos descritos neste artigo mostraram que a utilização de comitês compostos de classificadores baseados em redes neurais convolucionais permite obter desempenhos superiores aos relatados na literatura recente para a tarefa de detecção de linguagem ofensiva em textos publicados em mídias sociais.

Foi possível verificar que, na abordagem estudada, há tendência a um pequeno aumento da acurácia dos comitês à medida que aumenta o número de classificadores que os compõem; por outro lado, o aumento do número de classificadores está associado ao aumento linear do seu tempo de execução, da ordem de um minuto para cada novo classificador inserido. Cabe, portanto, investigar que combinação entre acurácia e tempo de execução oferece a melhor relação custo-benefício em função dos requisitos dos usuários. Em trabalhos futuros, sugere-se que além de identificar se

um texto contém linguagem ofensiva seja classificado o que tipo de ofensa contida. Os tipos de ofensas a considerar incluem racismo, sexismo, homofobia, xenofobia e intolerância religiosa.

Outra linha de investigação promissora para o problema em questão consiste em utilizar arquiteturas baseadas no pré-treinamento de modelos gerais da língua usando grandes corpus de texto não rotulado, como o BERT [14], que recentemente começam a se destacar em várias tarefas de PLN, incluindo a classificação de texto, contribuído para definir o novo estado da arte.

AGRADECIMENTOS

Ao CNPq, pela bolsa que possibilitou a realização deste trabalho; ao Prof. Dr. Juan Manuel Adán Coello, pela orientação durante o desenvolvimento do projeto; e a Caio Lima e Souza Della Torre Sanches por partilhar experiências e resultados obtidos em trabalhos anteriores.

REFERÊNCIAS

[1] KUMAR, R.; BHANODAI, G.; PAMULA, R.; CHENNURU, M. R. “TRAC-1 Shared Task on Aggression Identification: IIT (ISM) COLING’18”, in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, p. 58–65.

[2] LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis, Mining Text Data, Springer, pp. 415–463, 2012.

[3] COSTA NETO, A. D.; ADÁN COELLO, J. M. Redes Neurais Convolucionais Aplicadas à Análise de Sentimentos. In: XXII Encontro de Iniciação Científica da PUC-Campinas, 2017, Campinas.

[4] TREVISAN, V.; ADÁN COELLO, J. M. Redes Neurais Profundas Aplicadas à Detecção de Posicionamento em Redes Sociais. In: XXIII Encontro de Iniciação Científica e VIII Encontro de Iniciação em Desenvolvimento Tecnológico da PUC-Campinas, 2018, Campinas.

[5] PANG, B.; LEE, L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, v. 2, n. 1-2, p. 1–135, 2008.

[6] SOBHANI, P.; MOHAMMAD, S. M.; KIRITCHENKO, S. Detecting Stance in Tweets And Analyzing its Interaction with Sentiment. Proc. SemEval 2016. Disponível em <https://www.aclweb.org/anthology/S/S16/S16-2.pdf#page=177> Acesso em 29 jun. 2020.

[7] SANCHES, C. L. S. D. T.; ADÁN COELLO, J. M. Detecção de Posicionamento em Mídias Sociais Usando Comitês de Classificadores. In: Anais do XXIV Encontro de Iniciação Científica da PUC-Campinas. 2019 set 24-26; Campinas, São Paulo.

[8] GALLANT S. I. Neural network learning and expert systems. MIT press; 1993.

[9] BRITZ, D. Understanding Convolutional Neural Networks for NLP. Disponível em <http://www.wildml.com/2015/11/understandin g-convolutional-neural-networks-for-nlp/> Acesso em 02 jul. 2020.

[10] NOGUEIRA, C.; GATTI, M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. Disponível em <http://anthology.aclweb.org/C/C14/C14-1008.pdf> Acesso em 02 jul. 2020.

[11] FÜRNKRANZ, J. Ensemble Classifiers. [acesso em 20 jun. 2020]. Disponível em: https://www.ic.unicamp.br/ wainer/cursos/1s2012/mc906/ensembles.pdf

[12] DIETTERICH T. G. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg.

[13] PELLE, R; ALCÂNTARA, C; MOREIRA, V. P. “A Classifier Ensemble for Offensive Text Detection”, in Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, 2018, p.237-243.

[14] DEVLIN, M.-W. CHANG, K. LEE, e K. TOUTANOVA, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.