Classificação de Linguagem Tóxica com Modelos de Aprendizado Profundo

Thiago Severo Garbin
Porto Alegre, Brasil
tsgarbin@gmail.com

Henrique Schwab Gelatti
Porto Alegre, Brasil
schwabhenrique@gmail.com

***Resumo*—A utilização de redes neurais profundas pode ser de grande auxílio para a identificação de mensagens tóxicas. Neste trabalho se utilizou um conjunto de dados que continha tais mensagens para a comparação de oito diferentes arquiteturas de redes, envolvendo tanto camadas convolucionais quanto camadas recorrentes. Na análise das métricas utilizadas, verificou-se, em geral, um desempenho levemente superior das redes que continham apenas camadas convolucionais em detrimento daquelas com camadas recorrentes.**

***Palavras-Chave*—convolucional, recorrente, redes neurais, mensagem, tóxico**

I. Introdução

Há alguns anos, *Twitter* e *Facebook*, as principais redes sociais da atualidade, passaram a adotar medidas mais rígidas contra a disseminação de mensagens de ódio, em especial aquelas categorizadas como racistas ou homofóbicas. No entanto, há ainda diversos outros tipos de mensagens que podem ser classificadas como ofensivas a determinados indivíduos, grupos ou minorias, e que podem ter como alvo desde orientação religiosa a até mesmo características físicas. Desta forma, um desafio que se impõe é a obtenção de sistemas que tenham cada vez mais capacidade para detectar automaticamente as mensagens ofensivas e assim mitigá-las.

Neste sentido, no presente trabalho se propõe o treinamento e avaliação de modelos com diferentes arquiteturas de aprendizado profundo em um conjunto de dados que contém diversas espécies de mensagens classificadas como tóxicas chamado de *ToLD-BR (Toxic Language Dataset for Brazilian Portuguese)*, objetivando-se verificar por meio de diferentes métricas de avaliação as diferenças de desempenho das arquiteturas para o caso em questão.

II. Trabalhos Relacionados

A crescente preocupação para evitar a propagação de mensagens ofensivas também suscitou o surgimento de maior quantidade de trabalhos relacionados ao tema. No trabalho de Davidson et al. [1], os autores treinam um modelo com um dataset que contém 24.802 tweets que foram classificados entre três classes através de um sistema de *crowd-sourcing*, sendo as classes: linguagem de ódio, linguagem ofensiva e nenhum. Foram utilizadas técnicas como *steeming* e *TF-IDF* para preparação dos dados para posteriormente treinar diferentes modelos tradicionais e concluir que aquele com Regressão Logística apresentou melhor performance nas métricas analisadas. Entretanto, vale ressaltar que o dataset obtido pelos autores é bastante desbalanceado entre suas classes, pois enquanto a classe predominante corresponde a cerca de 77% do conjunto, a classe minoritária representa cerca de 6%.

Com o entendimento de que trabalhos anteriores não haviam explorado a questão do alvo dos comentários ofensivos, no trabalho de Zampieri et al. [2] é proposto um dataset para classificação hierárquica em três níveis de 14.100 mensagens do *Twitter*. Tais níveis consistem em: 1- Detecção da linguagem ofensiva, 2 - Categoria da linguagem ofensiva e 3 - Identificação do alvo da linguagem ofensiva. Para realizar as anotações também foi utilizado um sistema de *crowd-sourcing* e para modelagem foram testados modelos de aprendizado tradicionais e modelos de aprendizado profundo, em que o modelo configurado com *CNN* performou melhor que o configurado com *LSTM* bidirecional. A exemplo do dataset mencionado anteriormente, neste caso também há grande desbalanceamento, além de classes com poucas instâncias, em especial conforme se avança na hierarquia de classificação.

O dataset introduzido por Leite et al. [3], utiliza um contexto mais amplo, o da toxicidade, para englobar também mensagens consideradas apenas obscenas. Soma-se ao todo sete classes possíveis para as mensagens, considerando não tóxico, homofobia, obsceno, insulto, racismo, misoginia e xenofobia. Chamado de *ToLD-BR*, o dataset traz 21.000 mensagens em português brasileiro, também advindas do *Twitter* e classificadas manualmente por três voluntários cada. Os autores treinam modelos para classificação tanto com todas as classes como em uma versão binária do dataset, dando destaque para o modelo *BERT* pré treinado em multilínguas e apenas em português brasileiro, este último o qual obteve o melhor resultado na métrica *macro-F1*, a principal considerada, e que corresponde a um valor de 76% para o caso binário.

III. Planejamento

O presente trabalho se utilizou da versão binária do dataset *ToLD-BR*, a qual para ser obtida necessitou de pré-processamento próprio do dataset original, uma vez que este possuía grande desbalanceamento quando consideradas todas as sete classes. Desta forma, o dataset resultante ficou com duas classes, e balanceamento de 56% para a classe não tóxica e 44% para a classe tóxica. Dentro do cronograma proposto, esta foi a tarefa que se deu no primeiro dia, fazendo uso de

notebook do *Google Colaboratory* com a linguagem *Python* e as bibliotecas *Pandas* e *RE (Regular Expressions)*.

Nos quatro dias seguintes se deu a tarefa de implementação, em que foram utilizadas as bibliotecas *Tensor Flow* e *Keras* em notebook do *Google Colaboratory* com ambiente de execução em *GPU*, com vistas à aplicação de modelos de aprendizado profundo, além de bibliotecas auxiliares como *Numpy*, *Pandas* e *Plotly*.

IV. Implementação e Resultados

Em busca de maior preservação dos dados originais, realizou-se nas mensagens um pré-processamento próprio em que foram separadas as pontuações das palavras e eliminados os itens como citações a outros usuários e links. Posteriormente, na tokenização das palavras, optou-se por remover os filtros de pontuações para que estas também fossem tokenizadas, limitando-se, ao final, o tamanho do dicionário em 18.000 palavras.

Para a formação da arquitetura dos modelos, partiu-se de uma base que continha uma camada inicial de *embedding* com saída de 300 dimensões, e duas camadas finais, a de *pooling* máximo global de uma dimensão e a camada linear com ativação sigmóide. A função de custo selecionada foi a entropia cruzada binária e o otimzador escolhido foi o Adam.

Posteriormente, para o miolo dos modelos foram montadas oito diferentes combinações: a primeira apenas com uma camada de convolução 1D com 128 filtros, a segunda com uma camada de *dropout* seguida por uma de convolução 1D com 128 filtros, a terceira com uma camada de *dropout* seguida por uma de *LSTM* com 128 unidades, a quarta com uma camada de *dropout* seguida por uma de *LSTM* bidirecional com 128 unidades, a quinta com uma camada de *dropout* seguida por uma de *GRU* com 128 unidades, a sexta com uma camada de *dropout* seguida por uma de *GRU* bidirecional com 128 unidades, a sétima com uma camada de *dropout* seguida por uma de convolução 1D com 128 filtros e posteriormente uma de *LSTM* bidirecional com 128 unidades, e a oitava com uma camada de *dropout* seguida por uma de convolução 1D com 128 filtros e posteriormente uma de *GRU* bidirecional com 128 unidades.

Para o treinamento e a análise, o conjunto de dados foi separado em 80% para treino, 10% para avaliação e 10% para teste, bem como foram consideradas as métricas de acurácia, precisão, revocação e medida F (*F1 Score*). Após rodar cada modelo por 10 épocas e considerando a época em que cada um obteve seu menor valor de custo na validação, tem-se que, para a métrica de acurácia, o segundo modelo obteve o melhor valor no conjunto de validação e o primeiro modelo no conjunto de teste, com 77,43% e 76,95%, respectivamente. Para a métrica de precisão, o sexto modelo obteve o melhor valor no conjunto de validação e o sétimo modelo no conjunto de teste, com 73,57% e 73,37%, respectivamente. Para a métrica de revocação, o segundo modelo obteve o melhor valor no conjunto de validação e também no conjunto de teste, com 79,68% e 78,38%, respectivamente. Por fim, para a métrica de medida F, o segundo modelo novamente obteve o melhor valor no conjunto de validação e também no conjunto de teste, com 75,96% e 74,41%, respectivamente. Deste modo, percebe-se que os modelos apenas com camadas convolucionais obtiveram, no geral, resultados um pouco superiores aos modelos com camadas recorrentes.

V. Conclusão

Detectar potenciais mensagens tóxicas envolve algumas especificidades. Determinadas palavras podem ser consideradas tóxicas em um contexto, como quando se compara uma pessoa ou grupo a determinado animal, mas não tóxicas em outro contexto, como quando se está falando a respeito deste animal especificamente. Uma vez que o dataset utilizado tem originalmente a classe obsceno, a qual contém em alguns casos apenas palavras de baixo calão, isto pode fazer com que o modelo acabe por classificar elogios como tóxicos, caso contenham tais palavras. Outros fatores a serem considerados também são qualidade e o tamanho do dataset, uma vez que a classificação considerando a toxicidade de uma mensagem carrega consigo certa subjetividade e um maior conjunto de treinamentos tende a beneficiar o aprendizado do modelo.

No presente trabalho, foi possível verificar uma boa assertividade dos modelos de redes neurais profundas, em especial aqueles que continham apenas camadas convolucionais, pois obtiveram desempenho levemente superior aos modelos com camadas recorrentes. Para maior exploração dos resultados deste estudo e possível obtenção de melhores métricas, coloca-se a possibilidade futura de otimização de hiperparâmetros por meio de um grid search junto a uma validação cruzada.

Referências

[1] T. Davidson et al. "Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the 11th International AAAI Conference on Web and Social Media, 2017 . Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/14955. Acessado em: 04 ago. 2021.

[2] M. Zampieri et al. "Predicting the Type and Target of Offensive Posts in Social Media." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019. Disponível em: https://aclanthology.org/N19-1144.pdf. Acessado em: 04 ago. 2021.

[3] J. A. Leite et al. "Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis." Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020. Disponível em: https://aclanthology.org/2020.aacl-main.91.pdf. Acessado em: 04 ago. 2021.