Detectando linguagem ofensiva em tweets utilizando modelos Transformer

Offensive language detection using Transformer models

Marcos Aurélio Hermógenes Boriola (orientado) *, Gustavo Henrique Paetzold (orientador)†

RESUMO

Devido a facilidade de uso e grande número de acessos em plataformas de redes sociais é cada vez mais comum a presença de publicações que contenham algum tipo ataque, ameaça, ódio, palavras de baixo calão e afins, estes tipos de publicações são exemplos de textos com linguagem ofensiva. No intuito de controlar este tipo de conteúdo, sistemas de classificação são criados para detectar automaticamente linguagem ofensiva em textos e, no caso deste trabalho em específico, um sistema de classificação foi desenvolvido a partir de modelos baseados em Transformer. Os modelos foram refinados a partir de conjuntos de dados contendo tweets na língua inglesa rotulados como ofensivos (OFF) ou não ofensivos (NOT) e para avaliar o desempenho obtido as métricas Macro F1-score, Precisão e Revocação foram utilizadas. No total foram 11 modelos analisados onde o modelo de melhor performance superou os resultados obtidos pelo estado da arte na tarefa compartilhada OffensEval 2020.

Palavras-chave: Classificação de Texto. Processamento de Linguagem Natural. Redes Neurais Artificiais. Modelos Transformer.

ABSTRACT

Due to the ease of use and large number of accesses on social networking platforms, the presence of publications containing some type of attack, threat, hatred, profanity and related terms is increasingly common, these types of publications are examples of texts with offensive language. In order to control this type of content, classification systems are created to automatically detect offensive language in texts and, in this specific work, a classification system was developed from models based on Transformer. The models were refined from datasets containing tweets in English labeled as offensive (OFF) or non-offensive (NOT) and to evaluate the performance obtained the Macro F1-score, Accuracy and Recall metrics were used. In total, 11 models were analyzed where the model with the best performance surpassed the results obtained by the state of the art in the OffensEval 2020 shared task.

Keywords: Text Classification. Natural Language Processing. Artificial Neural Networks. Transformer Models.

1 INTRODUÇÃO

As plataformas online de redes sociais são um meio de facilitar a comunicação entre as pessoas, sejam elas amigos, parentes, com interesses em comum, e afins. O uso destas plataformas vai além de apenas possibilitar a comunicação entre pessoas distantes, elas também permitem que as pessoas compartilhem conteúdos de

* Engenharia de Computação; marcosboriola@alunos.utfpr.edu.br.
† Coordenação de Engenharia de Computação - Campus Toledo; ghpaetzold@utfpr.edu.br; https://orcid.org/0000-0001-9951-050X.

variados tipos onde os outros usuários podem interagir com estas publicações. Desde o surgimento das redes sociais, cada vez mais a utilização das mesmas vem crescendo, de forma que isto acaba atingindo mais pessoas com diferentes pensamentos, culturas e histórias (PERRIN, 2015). Com a facilidade de encontrar pessoas e conteúdos dentro destas plataformas, há aqueles usuários que as utilizam como meio de cometer ataques e tecer comentários negativos contra pessoas, grupos, instituições, gerando, assim, conteúdos ofensivos nessas plataformas (CHETTY; ALATHUR, 2018).

Tendo em vista este problema com linguagem ofensiva, termo que pode ser descrito como o tipo de linguagem que contém insultos, ameaças, ofensas, palavras de baixo calão, obscenidade e afins (ZAMPIERI et al., 2019; ROSENTHAL et al., 2020), plataformas de redes sociais tendem a trabalhar com formas de detectar automaticamente este tipo de conteúdo devido ao alto volume de dados gerados a cada instante, tornando impraticável a utilização de análise puramente humana para este fim. Então a pergunta que surge é a seguinte: como detectar linguagem ofensiva de forma ágil e eficaz?

Uma forma de fazer isso é através de técnicas utilizando Inteligência Artificial (IA). Por ser de conteúdo textual, a área de Processamento de Linguagem Natural (PLN) proporciona técnicas que são úteis e auxiliam na análise dos textos que contenham linguagem ofensiva. Dentro de PLN o problema aqui destacado é dito como uma Classificação Binária de Texto, onde análises são feitas com base nas características textuais e um algoritmo consegue definir se uma dada sentença possui ou não linguagem ofensiva.

Recentemente, uma classe de modelos surgiu em meio aos modelos de Redes Neurais Artificias (RNA), esta classe conhecida por Transformer teve seu primeiro artigo publicado no ano de 2017 (VASWANI et al., 2017) e desde então variações do modelo original vêm se tornando o estado da arte em diversas tarefas dentro e fora da área de PLN. Devido a robustez e a eficácia desta classe de modelos, este trabalho foca na utilização destes modelos na resolução do problema descrito.

Dado o número cada vez maior de usuários e ofensas sendo publicadas, alternativas para filtragem deste tipo de conteúdo se tornam cada vez mais relevantes e necessárias, assim sendo, o intuito deste trabalho se dá pela utilização de modelos do tipo Transformer alinhado a técnicas de PLN para detectar conteúdo ofensivo de forma automática, ágil e eficaz.

A ideia deste trabalho é baseada na tarefa compartilhada[1] OffensEval[2] que durante os anos de 2019 e 2020 reuniu diversas equipes para atacar o problema de detecção, categorização e identificação do alvo de linguagem ofensiva. Nesta tarefa compartilha são utilizados tweets em inglês como dado base e este trabalho utilizou destes mesmos dados focando na tarefa de detecção de linguagem ofensiva.

2 MÉTODO

Para o desenvolvimento de um sistema de classificação binária de texto baseado em RNA, neste caso em específico, para detecção de linguagem ofensiva em tweets são necessários três itens: conjunto de dados, modelo para classificação e métricas para avaliação de performance. Estes itens estão descritos nas subseções a seguir.

[1] Uma tarefa compartilhada tem como objetivo reunir as melhores soluções para um tema desafiador através da contribuição de muitas pessoas.
[2] https://sites.google.com/site/offensevalsharedtask/home

2.1 Conjuntos de Dados

Como mencionado, este trabalho se baseia na OffensEval, e portanto, os dados utilizados são os fornecidos por esta tarefa compartilhada. Os autores da OffensEval disponibilizam dois conjuntos de dados, o *Offensive Language Identification Dataset* (OLID) (ZAMPIERI et al., 2019) e o *Semi-Supervised Offensive Language Identification Datatset* (SOLID) (ROSENTHAL et al., 2020) que estão disponíveis no portal da competição[2].

O conjunto de dados OLID possui 14.100 tweets onde os mesmos foram rotulados manualmente por anotadores através de uma plataforma de crowd-sourcing. O conjunto de dados SOLID possui pouco mais de 9 milhões de tweets que foram anotados através de modelos de Machine Learning semi-supervisionados e o mesmo veio como uma evolução do conjunto anterior para que modelos de classificação mais robustos pudessem ser treinados.

Na Tab. 1 está descrito a quantidade de cada rótulo presente nos conjuntos de dados mencionados. O rótulo OFF é referente aos tweets considerados ofensivos e o rótulo NOT é referente aos tweets não ofensivos.

Tabela 1 – Distribuição dos rótulos nos conjuntos de dados.

Conjunto de Dados	Rótulo	Quantidade	Total
OLID	OFF	4.640	14.100
	NOT	9.460	
SOLID	OFF	1.448.861	9.089.140
	NOT	7.640.279	

Fonte: Adaptado. Zampieri et al. (2019) e Rosenthal et al. (2020).

Para melhor ilustrar os dados a serem classificados, na Tab. 2 há alguns exemplos de tweets presentes nos conjuntos de dados.

Tabela 2 – Exemplos de tweets nos conjuntos de dados.

Tweet	Rótulo
@USER He is so generous with his offers.	NOT
@USER She is a goddess	NOT
@USER Who the fuk are you	OFF
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF
@USER Figures! What is wrong with these idiots? [...]	OFF

Fonte: Adaptado. Zampieri et al. (2019) e Rosenthal et al. (2020).

Para treinamento dos modelos foram separados os conjuntos de dados a fim de comparar os modelos treinados com cada um dos conjuntos. O conjunto OLID foi utilizado em sua totalidade por possuir uma baixa quantidade de instâncias. No SOLID foram selecionados 50.000 instâncias, 25.000 de cada rótulo; por ser um conjunto de dados onde os rótulos são baseados em probabilidades, foram selecionados apenas tweets onde a probabilidade ficava entre 0,1 e 0,3 para o rótulo NOT e entre 0,7 e 0,9 para o rótulo OFF, além de selecionar apenas aqueles onde o desvio padrão fosse menor, isso tudo foi feito para evitar outliers. Além disso, para teste dos modelos treinados foi utilizado o conjunto de testes disponibilizado na OffensEval 2020, conjunto no qual possui 5.993 instâncias, sendo 3.002 instâncias do rótulo OFF e 2.991 do rótulo NOT.

2.2 Modelos

Para realização deste trabalho foram utilizados os seguintes modelos no qual todos são variações do modelo Transformer[3]: $\text{BERT}_{\text{BASE}}$-{CASED, UNCASED}; BERTweet e BERTweet-Offensive; BORT; DistilBERT-{CASE, UNCASED}; $\text{ELECTRA}_{\text{BASE}}$-Discriminator; RoBERTa; Twitter-RoBERTa e Twitter-RoBERTa-Offensive.

Um detalhe a ser mencionado sobre alguns dos modelos utilizados é que aqueles modelos que contêm "tweet"/"Twitter" e/ou "offensive" em seu nome foram pré-treinados com tweets comuns e/ou que contenham linguagem ofensiva, respectivamente.

Como todos estes modelos citados são pré-treinados, os mesmos foram refinados através de uma técnica de re-treinamento de modelo conhecida por *Fine-tuning*. A técnica *Fine-tuning* ajusta os parâmetros internos de um modelo já treinado com base em novos dados de entrada com o objetivo de dar ao modelo um aprendizado mais focado em um domínio específico.

Para uma melhor avaliação da generalização de cada modelo, a técnica de Validação Cruzada (SALKIND, 2010) foi empregada durante o treinamento. O conjunto de dados de treinamento foi divido em 10 partes, sendo 9 delas utilizadas para treinar o modelo e 1 para usar de validação durante o treinamento. Houve uma rotação das partes utilizadas para cada função a fim de que todas as partes fossem utilizadas tanto para treinamento quanto validação. Portanto, ao final do treinamento 10 modelos foram obtidos, cada um treinado por partes diferentes do conjunto de dados de treinamento. Para avaliar o desempenho do modelo como um todo, os resultados foram juntados através de uma média ponderada onde os modelos com maior pontuação Macro F_1-score (métrica descrita na próxima seção) tiveram maior peso.

2.3 Métricas de Avaliação

Um sistema de classificação analisa e processa o dado de entrada e dá como saída a classe/grupo que aquele dado pertence, portanto, as métricas de avaliação de um sistema de classificação mensuram quão bem o modelo de classificação se sai em predizer a classe correta. As métricas mais comuns para análise de sistemas de classificação são: Precisão, que informa a fração de predições corretas em relação a todas as predições de uma determinada classe; Revocação ou recall, que mostra a razão entre as predições verdadeiras em relação a todas as classes; e a F_1-score, que faz uma média harmônica entre as métricas anteriores pois as mesmas tendem a ser antagônicas. Para se obter uma única pontuação do sistema todo em relação a todas as classes classificadas é utilizada a métrica Macro F_1-score que calcula uma igual contribuição de todas as classes da métrica F_1-score. As equações Eq. (1), Eq. (2) e Eq. (3) representam as métricas Precisão, Revocação e F_1-score respectivamente, onde o termo *VP* significa verdadeiro positivo, *FP* falso positivo e *FN* falso negativo, e na Eq. (4) o termo *N* representa o número de classes do sistema.

$$P = \frac{VP}{VP + FP} \tag{1}$$

$$R = \frac{VP}{VP + FN} \tag{2}$$

[3] Detalhes sobre os modelos podem ser conferidos em: https://huggingface.co/models

$$F_1 = 2 * \frac{P * R}{P + R} \tag{3}$$

$$Macro - F_1 = \frac{1}{N} \sum_{i=1}^{N} F_{1i} \tag{4}$$

3 RESULTADOS

A Tabela 3 mostra os resultados obtidos tanto no treinamento dos modelos com o conjunto de dados OLID quanto com o SOLID, além disso contém também os resultados do pior caso do sistema de classificação, quando generaliza todos os dados como sendo de apenas um rótulo, esses resultados servem de base para uma comparação do desempenho dos modelos. As abreviações F_1, P e R presentes na tabela significam, respectivamente, as métricas Macro-F_1, Precisão e Revocação.

Tabela 3 – Resultados Obtidos.

Dados de Treinamento	Modelos	F_1	P	R
OLID	$BERT_{BASE}$-CASED	0,9217	0,9298	0,9220
	$BERT_{BASE}$-UNCASED	0,9216	0,9290	0,9218
	BERTweet-Offensive	0,9259	0,9337	0,9261
	BERTweet	0,9250	0,9335	0,9253
	BORT	0,9004	0,9030	0,9005
	DistilBERT-CASED	0,9192	0,9269	0,9195
	DistilBERT-UNCASED	0,9207	0,9280	0,9210
	$ELECTRA_{BASE}$-DISCRIMINATOR	0,9253	0,9328	0,9255
	$RoBERTa_{BASE}$	0,9235	0,9326	0,9238
	Twitter-RoBERTa-Offensive	0,9257	0,9312	0,9258
	Twitter-RoBERTa	0,9263	0,9324	0,9265
SOLID	BERTweet-Offensive	0,9215	0,9309	0,9218
	Twitter-RoBERTa-Offensive	0,9195	0,9287	0,9198
	Twitter-RoBERTa	0,9175	0,9275	0,9179
Linha de base	Tudo NOT	0,3329	0,2495	0,5
	Tudo OFF	0,3337	0,2505	0,5

Fonte: Autoria própria (2021).

Verificando os resultados demonstrados na Tab. 3 em relação ao OLID é possível verificar que a performance de todos os modelos foram bem semelhantes, com exceção do modelo BORT que possui uma discrepância maior em relação aos outros. Outro fato a ser observado é que mesmos os modelos que não haviam sido pré-treinados no domínio específico para este problema obtiveram um desempenho bem próximo dos modelos que tinham sido pré-treinados, assim, mostrando que generalizaram bem o novo conteúdo durante o refinamento dos mesmos.

Para o conjunto de dados SOLID foram apenas treinados os 3 melhores modelos obtidos no OLID e é possível ver que nos 3 os resultados são bem próximos. Os resultados do OLID, mesmo com um conjunto de dados menor possui vantagem no conjunto de testes.

Como o desafio deste trabalho já vem sendo tratado nas edições da OffensEval, na última edição (OffensEval 2020), a equipe que ficou na primeira colocação obteve uma pontuação Macro F_1-score de 0,9204 pontos e com isso se tornou o estado da arte para esta tarefa. Saber do melhor resultado obtido facilita a comparação dos

resultados obtidos neste trabalho com o estado da arte.

4 CONCLUSÕES

Neste trabalho foi apresentado o desempenho de alguns modelos Transformer na tarefa de classificar tweets em inglês como ofensivos ou não ofensivos e é possível observar nos resultados obtidos que muitos modelos obtiveram resultados melhores que o estado da arte. As melhores performances foram obtidas nos modelos treinados com o OLID, que é um conjunto de dados menor e com desbalanceamento nas classes. Isso mostra que a quantidade de dados para refinamento dos modelos não foi um fator decisivo na generalização dos modelos, mas sim a qualidade da anotação dos dados, que no caso do OLID foram anotados manualmente.

Muitas análises ainda podem ser feitas em cima desta tarefa em trabalhos futuros como: fazer análises aprofundadas nas sentenças que sofreram erros de predição; agrupar os melhores modelos atuais e verificar o desempenho deste sistema maior; preparar um conjunto de dados onde há a separação de tweets realmente ofensivos daqueles que somente utilizam palavras de baixo calão como intensificadores de frase.

AGRADECIMENTOS

Meus agradecimentos ao meu orientador Gustavo H. Paetzold por todo auxílio durante a realização dos trabalhos e por ter me indicado este desafio, e também gostaria de agradecer a Fundação Araucária pelo apoio através do financiamento para realização deste projeto de iniciação científica.

REFERÊNCIAS

CHETTY, Naganna; ALATHUR, Sreejith. Hate speech review in the context of online social networks. **Aggression and Violent Behavior**, v. 40, p. 108–118, 2018. ISSN 1359-1789.

PERRIN, Andrew. Social Networking Usage: 2005-2015. **Pew research center**, out. 2015. Disponível em: .

ROSENTHAL, Sara et al. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification, 2020.

SALKIND, Neil J. **Encyclopedia of Research Design**. 1. ed. [S.l.]: SAGE Publications, 2010. v. 1, p. 314–316. ISBN 9781412961271.

VASWANI, Ashish et al. Attention is All You Need. In: PROCEEDINGS of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

ZAMPIERI, Marcos et al. Predicting the Type and Target of Offensive Posts in Social Media. In: PROCEEDINGS of NAACL. [S.l.: s.n.], 2019.