DETECÇÃO AUTOMÁTICA DE LINGUAGEM OFENSIVA E/OU DE ÓDIO EM LÍNGUA PORTUGUESA

Gabriel Gonçalves de Matos (PIC/CNPq/FA/UEM), Valéria Delisandra Feltrim (Orientadora), E-mail: ra116843@uem.br

Universidade Estadual de Maringá / Centro de Tecnologia / Maringá, PR.

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: Inteligência Artificial; Processamento de linguagem natural; Linguagem ofensiva.

RESUMO

A web e as redes sociais são espaços de comunicação que permitem a interação, a participação e a mobilização dos usuários, mas também podem ser usados para disseminar conteúdos prejudiciais, como ódio e intolerância. Em vista disso, torna-se necessário buscar formas de combater tal ação de modo a amenizar os danos sociais causados pela propagação desse tipo de conteúdo. Nesse sentido, diversos modelos para a detecção automática de discurso de ódio têm sido propostos pela comunidade científica, inclusive para a língua portuguesa. Tais modelos são treinados por meio de algoritmos de aprendizagem de máquina supervisionados e ajustados para uma base de dados particular. Nesse contexto, o objetivo deste projeto foi avaliar diferentes modelos classificadores de linguagem tóxica e/ou de ódio, visando identificar configurações de aprendizado que se destaquem em termos de desempenho em diferentes bases. Também se verificou a variação de desempenho dos modelos aprendidos quando avaliados em bases de dados diferentes das utilizadas no treinamento. Os resultados mostraram que os modelos BERT superaram os demais e que, dependendo da base de treinamento, há uma queda de desempenho quando o modelo é avaliado com comentários provindos de uma base diferente daquela utilizada para o treinamento.

INTRODUÇÃO

As redes sociais se tornaram um espaço de comunicação, informação e entretenimento para milhões de pessoas no mundo. No entanto, esse ambiente também é palco de manifestações de ódio e intolerância contra grupos e indivíduos que são vistos como diferentes. Diante dessa situação surge a necessidade de desenvolver mecanismos que possam identificar comentários ofensivos presentes na Web. A comunidade científica e, em particular, pesquisadores da área de Processamento de Linguagem Natural (PLN) têm envidado esforços na construção

de sistemas computacionais capazes de detectar esse tipo de conteúdo automaticamente (AL-HASSAN e AL-DOSSARI, 2021; CORAZZA et al., 2020). Nessa mesma direção, diversos estudos têm disponibilizado bases de dados sobre linguagem ofensiva e/ou de ódio em diversas línguas, comumente coletadas a partir de redes sociais, bem como têm proposto modelos baseados em aprendizado de máquina (AM) para a detecção desse tipo de conteúdo. Embora a maior parte da literatura nessa área seja voltada para a língua inglesa, o número de trabalhos relacionados à detecção de linguagem ofensiva e/ou de ódio em língua portuguesa também tem crescido (LEITE et al., 2020; VARGAS et al., 2022). Dado esse contexto, o objetivo deste trabalho foi avaliar diferentes modelos de AM para a detecção de linguagem tóxica e/ou de ódio, visando identificar configurações de aprendizado que se destaquem em termos de desempenho com diferentes bases de dados. Também se verificou a variação de desempenho dos modelos aprendidos quando avaliados em bases de dados diferentes daquela utilizada no treinamento.

MATERIAIS E MÉTODOS

Foram utilizadas duas bases de dados neste projeto: a ToLD-Br (*Toxic Language Dataset for Brazilian Portuguese*) (LEITE et al., 2020) e a HateBR (VARGAS et al., 2022). A primeira é composta por comentários coletados da rede social Twitter(X) entre agosto e julho de 2019, sendo 11.745 rotulados como não tóxico (0) e 9.255 como tóxico (1), totalizando 21.000 tweets. A segunda é constituída por comentários retirados da rede social Instagram entre agosto de 2020 e janeiro de 2021, contendo 7.000 comentários, sendo 3.500 não ofensivos (0) e 3.500 ofensivos (1).
As representações, métodos e algoritmos de AM utilizados neste projeto também se basearam nos trabalhos de Leite et al. (2020) e Vargas et al. (2022), com algumas adaptações nas ferramentas e recursos utilizados.
Os modelos de classificação foram construídos utilizando a linguagem Python e tal processo pode ser dividido nas etapas descritas a seguir.

Pré-Processamento

Os classificadores foram construídos com e sem pré-processamento das bases de dados. O primeiro passo do pré-processamento dos comentários consistiu em deixar todas as letras minúsculas, a fim de reduzir a complexidade e a viabilidade dos dados, facilitando o aprendizado do modelo. O próximo passo foi realizar a tokenização, que consiste em dividir os textos em unidades indivisíveis, chamadas tokens. Para isso foi utilizada a biblioteca spaCy. A SpaCy também foi utilizada para a remoção de *stop words*, que são palavras muito comuns e pouco informativas, e sinais de pontuação. Além disso, foi feita a lematização dos tokens dos comentários. O lema é a forma básica e canônica de uma palavra. Por exemplo, “cantando”, “cantou” e “cantaria” tem o lema “cantar”. A lematização pode ser útil para simplificar e padronizar os textos, reduzindo a variação morfológica e facilitando a análise semântica.

Classificação dos comentários

Foram utilizados cinco algoritmos de classificação nos experimentos, a saber: Naive Bayes (NB), Support Vector Machine (SVM), Multilayer Perceptron(MLP), Regressão Logística(LR) e o BERT. Os modelos BERT foram obtidos a partir do *fine-tuning* do modelo BERTimbau (SOUZA et al., 2020), que é um modelo BERT pré-treinado para português brasileiro. Devido às limitações de hardware, foi utilizado o BERT-Base.
Em um primeiro experimento, os modelos foram treinados e avaliados com as bases ToLD-Br e HateBR, com e sem o pré-processamento dos dados. Cada base foi dividida em partições de treino e teste na proporção 80%-20% e todos os classificadores foram treinados e avaliados utilizando as mesmas partições.
Em um segundo experimento, um classificador BERT treinado com 100% da base ToLD-Br foi avaliado utilizando os comentários da HateBR. De forma similar, um classificador BERT treinado com 100% da HateBR foi avaliado utilizando a ToLD-Br.
Para realizar os treinamentos e validações dos modelos foram utilizadas as bibliotecas de aprendizado de máquina scikit-learn e Simple Transformer.

RESULTADOS E DISCUSSÃO

A Tabela 1 mostra os melhores resultados obtidos no primeiro experimento, em termos da métrica micro-F1. Os valores sinalizados com asterisco (*) foram obtidos com pré-processamento dos dados, enquanto os demais foram obtidos sem pré-processamento.

Tabela 1. Micro-F1 dos classificadores avaliados.

Bases	NB	SVM	MLP	LR	BERT
ToLD-BR	0,70*	0,75	0,70*	0,75	**0,78**
HateBR	0,85*	0,85*	0,85*	0,85*	**0,92**

Fonte: Autor.

É possível observar que os modelos BERT, sem pré-processamento, obtiveram desempenho superior aos demais. No caso do classificador BERT-ToLD-BR, os resultados foram similares ao melhor resultado apresentado por Leite et al. (2020) (0,76 de micro-F1 com um classificador BERT). Já os resultados do BERT-HateBr foram superiores ao melhor resultado apresentado por Vargas et al. (2022) (0,85 de micro-F1 com um classificador SVM e tf-idf). Cabe destacar que Vargas et al. (2022) não avaliaram classificadores baseados no BERT.
No segundo experimento, foram treinados modelos BERT utilizando 100% das bases ToLD-BR e HateBR, chamados de BERT-ToLD-BR-Full e BERT-HateBR-Full. O classificador BERT-ToLD-BR-Full foi avaliado com a base HateBR e obteve 0,78 de micro-F1. O classificador BERT-HateBR-Full, por sua vez, foi avaliado com a

base ToLD-BR e obteve 0,71 de micro-F1. Quando comparados aos resultados do primeiro experimento, é possível observar que o BERT-ToLD-BR-Full manteve o desempenho, enquanto o BERT-HateBR-Full teve uma queda de 0,21 em pontos no desempenho quando avaliado com uma base diferente da utilizada no treinamento.

CONCLUSÕES

O objetivo deste estudo foi explorar diferentes modelos classificadores de linguagem ofensiva e/ou de ódio na língua portuguesa, bem como avaliar se há uma configuração que se destaque em termos de desempenho independente da base utilizada. Com base nos trabalhos de Leite et al. (2020) e Vargas et al. (2022), foram elaborados modelos para realizar a classificação de comentários. Os melhores modelos propostos utilizaram o modelo BERT sem a utilização do pré-processamento. O modelo BERT treinado com a base de Leite et al. (2020) manteve o desempenho em diferentes bases. O mesmo não foi observado para o modelo BERT treinado com a base de Vargas et al. (2022). Tal comportamento pode ser explicado pelas diferenças de tamanho e generalidade das bases, mas um estudo mais aprofundado se faz necessário para avaliar o comportamento dos classificadores e as diferenças das bases de dados utilizadas.

REFERÊNCIAS

AL-HASSAN, A.; AL-DOSSARI, H. Detection of hate speech in arabic tweets using deep learning. **Multimedia Systems**, 2021.

CORAZZA, M.; MENINI, S.; CABRIO, E.; TONELLI, S.; VILLATA, S. A multilingual evaluation for online hate speech detection. **ACM Trans. Internet Technol.**, v. 20, n. 2, 2020.

LEITE, J.A.; SILVA, D.F.; BONTCHEVA, K.; SCARTON, C. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In: 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020. **Proceedings…** Association for Computational Linguistics, 2020, p. 914–924.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems (BRACIS), 2020. **Proceedings...** Springer, Cham, 2020. p. 403–417.

VARGAS, F.; CARVALHO, I.; GÓES, F.; PARDO, T.A.S.; BENEVENUTO, F. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In: 13th Conference on Language Resources and Evaluation (LREC 2022), 2022, Marseille. **Proceedings…** European Language Resources Association (ELRA), 2022. p. 7174–7183.