Análise sentimental em português de textos extraídos do Twitter com modelos BERT

Sentimental analysis in Portuguese of texts extracted from Twitter with BERT models

Vinícius Pinheiro Winter *, Gustavo H. Paetzold (orientador)†

RESUMO

Grande parte dos dados disponíveis na internet são fornecidos por meio da escrita, meio onde as áreas de pesquisa têm expandido muito nos últimos anos, principalmente com o uso e desenvolvimento de modelos de interpretação textual, como por exemplo os modelos BERT. Esta técnica foi um marco para o processamento de linguagem natural e diversas aplicações surgiram desde então. Umas destas aplicações que se aperfeiçoaram com o surgimento da técnica, foi a análise sentimental com base em texto. Neste projeto propõe-se o uso do BERT para desenvolvimento de um modelo de análise de sentimentos baseado em textos extraídos de redes sociais, para identificar casos de sentimentos negativos em publicações. O estudo em questão será aplicado à língua Portuguesa devido à escassez de material nesta língua, visto que a maioria das aplicações é feita em Inglês. O presente documento pretende validar o modelo desenvolvido com a finalidade de definir a polaridade dos sentimentos e servir de material de estudo para uma possível aplicação na identificação de traços de comportamento preocupantes através de análise em redes sociais.

Palavras-chave: Redes Sociais. Saúde Mental. Twitter. Suicídio.

ABSTRACT

Much of the data available on the internet is provided through writing, where research areas have expanded a lot in recent years, especially with the use and development of textual interpretation models, such as the BERT models (DEVLIN, 2019) . This technique was a milestone for natural language processing and several applications have emerged since then. One of these applications that improved with the advent of the technique was text-based sentimental analysis. In this project, the use of BERT is proposed to develop a feeling analysis model based on texts extracted from social networks, to identify cases of negative feelings in publications. The study in question will be applied to the Portuguese language due to the scarcity of material in this language, since most applications are made in English. This document manages to assemble a model that can define the polarity of feelings and serve as study material for a possible application in the identification of disturbing behavioral traits through analysis in social networks.

Keywords: Social Network. Mental Health. Twitter. Suicide.

1 INTRODUÇÃO

A análise de sentimentos é uma tarefa muito abordada no ramo do Processamento de Língua Natural (PLN), onde sua classificação pode ser feita através da polaridade do sentimento existe em um texto, sendo positivo, negativo e neutro (ROSA, 2015), ou ainda classificado por rótulos mais específicos como alegria, tristeza, raiva e surpresa (BRUM, 2015).

* Engenharia de Computação; winterv490@gmail.com.
† Professor Engenharia de Computação; ghpaetzold@outlook.com; https://orcid.org/0000-0001-9951-050X.

Esta classificação por sua vez, tem dois métodos de execução, o primeiro é realizado através de técnicas supervisionadas, que são baseadas em aprendizado de máquina e que necessitam de uma grande quantidade de dados rotulados para treinamento e teste. O segundo refere-se a técnicas não-supervisionadas, que utilizam tratamentos léxicos, cálculos e dicionários léxicos para classificação do sentimento contido em cada palavra do texto (ARAÚJO; GONÇALVES; BENEVENUTO, 2013).

Em resumo, no presente trabalho, iremos mostrar o motivo da escolha do modelo pré-treinado citado acima, o BERT, mostrar a dificuldade e problemas encontrados ao extrair dados das redes sociais bem como conseguir classificá-las e por fim os resultados obtidos durante a classificação de um outro dataset disponibilizado por um pesquisador envolvido em projetos de classificação de sentimentos. Também mostraremos ao fim do trabalho, sugestões e possíveis trabalhos que possam ser desenvolvidos utilizando os resultados obtidos no presente.

Conseguindo então responder a pergunta, é possível se desenvolver um modelo de classificação baseado no modelo BERT para classificação de sentimentos na Língua Portuguesa?

2 MATERIAIS E MÉTODOS

O trabalho em questão foi desenvolvido da seguinte forma, primeiro com a aquisição dos dados de treinamento, onde foi utilizada uma base de dados retirada do twitter, a partir de agora nos referenciaremos a esta como dataset de treinamento (GO; BHAYANI; HUANG, 2009), realizado o fine-tunning do modelo BERT disponibilizado pela neuralmind (SOUZA; NOGUEIRA; LOTUFO, 2020b), responsável pelo desenvolvimento do modelo destinado à Língua Portuguesa. Após este processo, foi encontrada outra base de dados, disponibilizada por (BRUM; NUNES, 2018), na qual a partir de agora chamaremos de dataset de teste, onde a base foi toda classificada através da disponibilização dos dados onde os mesmo poderiam ser classificados como positivo, neutro, negativo e nulo (quando não se conseguia definir nenhuma das classificações anteriores). Porém, para o presente trabalho o objetivo é a classificação binária da publicação, ou seja, positivo e negativo.

2.1 Modelo pré-treinado

Os dados citados abaixo foram retirados do artigo oficial publicado por (SOUZA; NOGUEIRA; LOTUFO, 2020a) a respeito do modelo BERT desenvolvido para análise de língua portuguesa nomeado como BERTimbau, onde precisou-se redefinir os principais meios do vocabulário do modelo devido a sua nova aplicação.

Para o pré-treinamento de dados deste modelo, foi usado o corpus brWaC (WAGNER FILHO et al., 2018) (Brazilian Web as Corpus), um crawl de páginas da web brasileiras que contém 2,68 bilhões de tokens de 3,53 milhões de documentos e é o maior corpus aberto em português até hoje. Além de seu tamanho, o brWaC é composto por documentos inteiros e sua metodologia garante alta diversidade de domínio e qualidade de conteúdo, características desejáveis para o pré-treinamento de BERT.

Em resumo o BERT é uma arquitetura de rede neural que pode ser treinada e utilizada em diversas tarefas, ou seja, o conjunto de dados do modelo é treinado em um corpus de texto (como o Wikipédia) e pode ser usado para desenvolver sistemas sem sair do zero, sua principal função é a transferência de conhecimento, ou seja, por meio de um treinamento prévio, a interpretação de texto pode ser repassada para diversas aplicações, como a classificação de sentimentos, sendo este um dos motivos pelo qual usaremos o modelo.

2.2 Dataset de treinamento

O primeiro objetivo era conseguir extrair dados diretamente do Twitter e utilizar algum modelo existente na literatura para criação do dataset de treinamento, pois assim seria possível extrair os dados de treinamento e validação dos mesmos usuários, mantendo assim um padrão na linguagem analisada pelo modelo. Porém apesar da existência de uma API do Twitter que facilita a extração deste tipo de informação, assim como também permitir a livre utilização meio a prévio cadastro em sua plataforma destinada a desenvolvedores, a API tem uma série de limitações quanto à número de requisições por mês e mais que isso, a cada período de tempo.

O dataset de teste foi extraído do Twitter e classificado utilizando o modelo disponibilizado por (GO; BHAYANI; HUANG, 2009). Os dados foram separados em dois grupos, um onde o assunto do Tweet era exclusivamente política e o outro onde o assunto não era filtrado, obtendo assim dados bem distintos e envolvendo diversos temas. Para o treinamento do modelo foram misturados os dois montantes obtidos e usados em conjunto para um melhor resultado. A distribuição das classificações do dataset de testes pode ser conferida na tabela 1.

Tabela 1 – Divisão de classificação dataset de treinamento

Classificação	Quantidade	Percentual %
Positiva	289327	34.61
Negativa	546739	65.39
Total	**836066**	**100**

Fonte: Autoria própria (2021).

2.3 Dataset de teste

Para os datasets da validação da classificação obtida pelo modelo, foram utilizados dois conjuntos de dados distintos, o primeiro, Base de ratings e avaliações do IMDb (RATINGS. . . , s.d.), um dataset baseado em classificações de filmes, onde os comentários da plataforma foram extraídos e classificados somente em positivos e negativos. O resumo de balanceamento nas classificações pode ser conferido na tabela 2.

Tabela 2 – Divisão de classificação dataset de teste - IMDb

Classificação	Quantidade	Percentual %
Positiva	24522	50
Negativa	24522	50
Total	**49044**	**100**

Fonte: Autoria própria (2021).

O segundo dataset de teste foi fornecido por (BRUM; NUNES, 2018) e é composto por 15000 tweets que foram classificados em positivo, negativo e neutro, baseado em uma descrição para cada um dos sentimentos, que foi seguido pelos responsáveis pela descrição, a discriminação das quantidades de dados classificados com cada uma das categorias pode ser conferida na tabela 3.

Para o presente trabalho, como o modelo foi treinado somente com as classificações positivas e negativas, iremos usar somente os 2/3 do dataset referentes a classificação de interesse disponibilizados, as proporções podem ser conferidas na tabela 3.

Tabela 3 – Divisão de classificação dataset de teste - Twitter

Classificação	Quantidade	Percentual %
Positiva	6648	62.87
Negativa	3926	37.13
Total	**10574**	**100**

Fonte: Autoria própria (2021).

3 RESULTADOS

Podemos verificar na tabela 4 comparativa entre os resultados obtidos no presente trabalho (identificados abaixo da linha divisória) e os obtidos em um trabalho de (CARDOSO; FERNANDES; AGUIAR, 2019), onde o objetivo era comparar a eficácia da classificação de sentimento de cada um dos algoritmos citados, bem como o de um comitê realizado juntando os algoritmos citados a fim de se obter um melhor resultado.

Os resultados abaixo da linha, são os obtidos através da classificação do modelo BERT e um comparativo dos resultados obtidos pelos modelos de Regressão Linear e Naive Bayes aplicado ao dataset de treinamento, para simples obtenção de comparativos.

Tabela 4 – Autoria prórpria

Algoritmos	Acurácia	Precisão	Recall	F1 Score	Erro
Naive Bayes	0.81	0.81	0.81	0.81	0.27
SVM	0.84	0.84	0.84	0.84	0.212
Árvore de Decisão	0.80	0.82	0.80	0.81	0.28
Random Forest	0.85	0.86	0.85	0.85	0.20
Regressão Logística	0.84	0.84	0.84	0.84	0.20
Comitê	**0.86**	**0.86**	**0.86**	**0.86**	**0.184**
Regressão Logística	**0.77**	0.78	0.70	**0.72**	-
Naive Bayes	**0.77**	0.78	0.70	**0.72**	-
BERT-IMDb	0.67	0.62	**0.86**	**0.72**	-
BERT-Twitter	0.58	**0.80**	0.39	0.53	-

Fonte: Autoria própria (2021).

Na parte de cima da tabela podemos perceber que o uso dos algoritmos em conjunto, gerando o comitê citado, resultou num resultado mais efetivo da classificação dos dados disponibilizados. Olhando para a parte de baixo da tabela, os dados gerados neste trabalho, percebemos que os modelos BERT obtiveram melhores resultados quando comparados a algoritmos de modo de treino supervisionado, porém, inferiores quando comparados ao obtidos pelo trabalho de (CARDOSO; FERNANDES; AGUIAR, 2019).

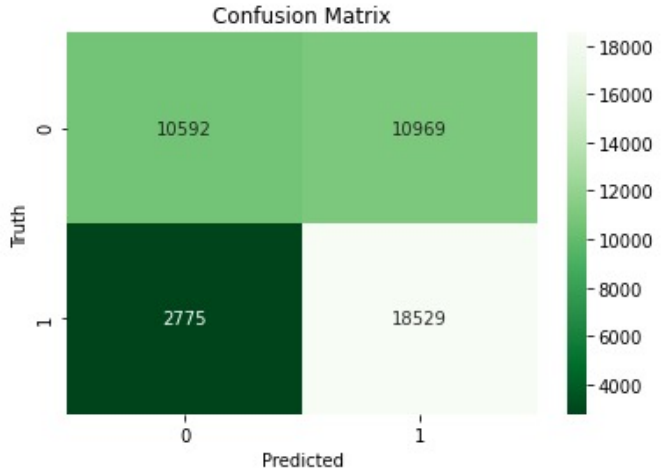
Na figura 1 podemos verificar a matriz de confusão obtida após a classificação da base de teste do IMDb (RATINGS. . . , s.d.) utilizando o modelo BERT treinado neste trabalho.

Percebemos aqui uma maior concentração de falso-positivo.

Em sequência podemos ver na figura 2, a matriz de confusão obtida após classificação dos da base de dados de teste do Twitter, utilizando o modelo BERT treinado no presente trabalho:

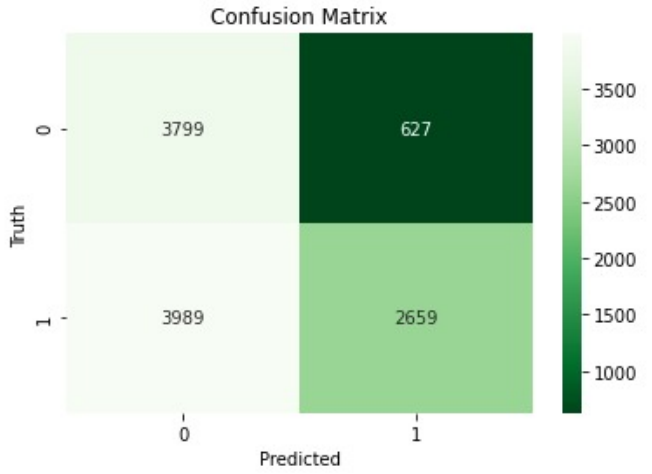
Percebemos aqui uma maior concentração de erro no falso-negativo.

Figura 1 – Matriz de confusão - IMDb



Fonte: Autoria própria (2021).

Figura 2 – Matriz de confusão - Twitter



Fonte: Autoria própria (2021).

4 CONCLUSÕES

A não assertividade dos resultados apresentados na seção 3 pode estar relacionado a falta de padrão nos dados de treinamento e classificação, devido a sua divergência em assuntos e linguagem usada em cada um dos tipos de dados fornecidos. Porém, percebe-se ainda, que os resultados são positivos, e que ao se realizar um trabalho de fine-tunning mais aprimorado, se pode obter resultados mais eficazes.

Conseguimos então mostrar que o modelo pre-treinado BERT, pode ser treinado e destinado à classificação de sentimento na Língua Portuguesa e então apresentar uma boa efetividade na aplicação da função de classificação de sentimento, podendo ser destinada a diversas outras finalidades dependentes desta classificação. Dentre as recomendações, está um possível trabalho de classificação mais refinada, não só destinada a polaridade do sentimento mas sim a classificação detalhada do mesmo como felicidade, raiva, tristeza, etc. . Podendo também ser destinada ao desenvolvimento de projetos mais ambiciosos, como o desenvolvimento de um estudo de variações de sentimento de usuários ao longo do tempo a fim de descobrir possíveis indícios a tendências de risco a vida.

AGRADECIMENTOS

Agradecimento especial a UTFPR que tornou possível fornecendo toda a estrutura necessária, também ao meu Orientador Dr. Gustavo H. Paetzol ao auxílio prestado durante o desenvolvimento de todas as partes do trabalho, assim como a família e amigos pelo apoio e incentivo durante as dificuldades encontradas neste período.

REFERÊNCIAS

ARAÚJO, Matheus; GONÇALVES, Pollyanna; BENEVENUTO, Fabrício. Measuring sentiments in online social networks. In: ANAIS do XIX Simpósio Brasileiro de Sistemas Multimídia e Web. Salvador: SBC, 2013. P. 97–104. Disponível em: .

BRUM, Henrico Bertini. **Análise de sentimentos para o português usando redes neurais recursivas**. 2015. Universidade Federal do Pampa.

BRUM, Henrico Bertini; NUNES, Maria das Graças Volpe. Building a sentiment corpus of tweets in brazilian portuguese. In: INTERNATIONAL Conference on Language Resources and Evaluation - LREC. [S.l.]: European Language Resources Association, 2018.

CARDOSO, Matheus; FERNANDES, Anita; AGUIAR, Sandro de. Análise de Sentimentos: Uma Comparação entre Diferentes Abordagens no Contexto da Língua Portuguesa. In: ANAIS da X Escola Regional de Informática de Mato Grosso. Cuiabá: SBC, 2019. P. 118–120. DOI: 10.5753/eri-mt.2019.8606. Disponível em: .

DEVLIN, J. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. [S.l.: s.n.], 2019.

GO, Alec; BHAYANI, Richa; HUANG, Lei. Twitter sentiment classification using distant supervision. **Processing**, v. 150, jan. 2009.

RATINGS, Reviews, and Where to Watch the Best Movies TV Shows. [S.l.]: IMDb.com. Disponível em: .

ROSA, Renata Lopes. **Análise de sentimentos e afetividade de textos extraídos das redes sociais**. 2015. Tese (Doutorado) – Escola Politécnica da Universidade de São Paulo.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: [s.l.: s.n.], out. 2020. P. 403–417. ISBN 978-3-030-61376-1. DOI: 10.1007/978-3-030-61377-8_28.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.], 2020.

WAGNER FILHO, Jorge A. et al. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In: PROCEEDINGS of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), mai. 2018. Disponível em: .