UTILIZANDO BERT PARA ANÁLISE DE SENTIMENTO EM CORPORA DE LÍNGUA PORTUGUESA

GUILHERME DA SILVA CAMARGO[1]; LARISSA ASTROGILDO DE FREITAS[2]

[1]*Universidade Federal de Pelotas – gdscamargo@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*

1. INTRODUÇÃO

Com o crescente uso das redes sociais enormes quantidades de dados são gerados diariamente, com isso aumenta a demanda de processamento deste tipo de dado. Diante desta demanda, surge a necessidade de se criar maneiras eficazes de processar esses dados. Isso pode ser feito através da área de Processamento de Língua Natural (PLN), que é uma área no qual se estuda maneiras de como o computador pode ser utilizado para entender e manipular textos ou áudios em linguagem natural (CHOWDHURY, 2003).

Dentro da área de PLN, existe a tarefa de Análise de Sentimento (AS), que é o campo de estudo que analisa as opiniões, avaliações, atitudes e sentimentos através de entidades e seus atributos expressos no texto escrito (LIU,2015). Para tal tarefa, existem duas abordagens principais: uma baseada em Aprendizado de Máquina (AM), onde um determinado algoritmo é treinado com dados já marcados, e outra baseada em léxico, que faz uso de uma lista pré-definida de palavras, onde cada palavra é associada com um sentimento específico (GONÇALVES, 2013). No presente trabalho, será utilizado uma abordagem baseada em AM, mais especificamente, utilizando um modelo de rede neural baseado em *transformers*, chamado BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN, 2018), em *datasets* de língua portuguesa.

BERT consiste num modelo pré-treinado, em grandes quantidades de dados não anotados, com duas tarefas não-supervisionadas: a primeira é a *masked language model* (MLM), onde uma porcentagem de cada texto é mascarada e então o modelo tenta prever tais palavras mascaradas; e a segunda se chama *next sentence prediction* (NSP), onde o modelo é alimentado com pares de sentenças e aprende a prever se uma é subsequente a outra (DEVLIN, 2018).

2. METODOLOGIA

No presente trabalho, inicialmente, será utilizado o corpora de língua portuguesa TweetSentBr (BRUM, 2017), que consiste em 15.000 *tweets* sobre shows de TV exibidos no primeiro semestres de 2017. Estes *tweets* foram classificados em três polaridades (positiva, negativa e neutra) em nível de documento por 7 anotadores.

BERT pode ser treinado com diferentes parâmetros e com textos de diferentes línguas, inicialmente, utilizamos o modelo pré-treinado BERT-base para múltiplos idiomas (DEVLIN, 2018), este modelo para múltiplos idiomas dá suporte para até 104 idiomas, com diferenciação entre maiúsculas e minúsculas.

No *fine-tuning*, onde o modelo é alimentado com o *dataset* a ser testado e as anotações servem para conferir as predições, foram realizados dois experimentos iniciais com um algoritmo de *fine-tuning* para classificação de texto proposto por (LI, 2020), onde as classes do algoritmo foram substituídas pelas classes do

corpora (positiva, negativa e neutra). No primeiro experimento foram utilizados os textos do corpora em seu estado original, em português, e no segundo experimento foram utilizados os textos traduzidos pelo Google Tradutor, em inglês.

3. RESULTADOS E DISCUSSÃO

Para fins comparativos, utilizaremos apenas a métrica Medida F, que é uma média harmônica entre a Precisão (proporção de positivos identificados corretamente) e Revocação (proporção de positivos identificados dentro todos).

As Tabelas 1 e 2 mostram a Medida F obtida para cada um dos dois experimentos realizados, ambos com 4 Épocas (do inglês *Epoch*).

Epoch	***Training Loss***	***Validation Loss***	**Medida F (%)**
1	0,91	0,78	64,87
2	0,78	0,94	67,25
3	**0,73**	**1,03**	**68,36**
4	0,64	1,35	67,79

Tabela 1: Experimento 1, textos de entrada em português.

Epoch	***Training Loss***	***Validation Loss***	**Medida F (%)**
1	0,90	0,82	66,01
2	0,79	0,92	65,97
3	**0,72**	**1,21**	**67,44**
4	0,64	1,41	66,78

Tabela 2: Experimento 2, textos de entrada traduzidos para inglês.

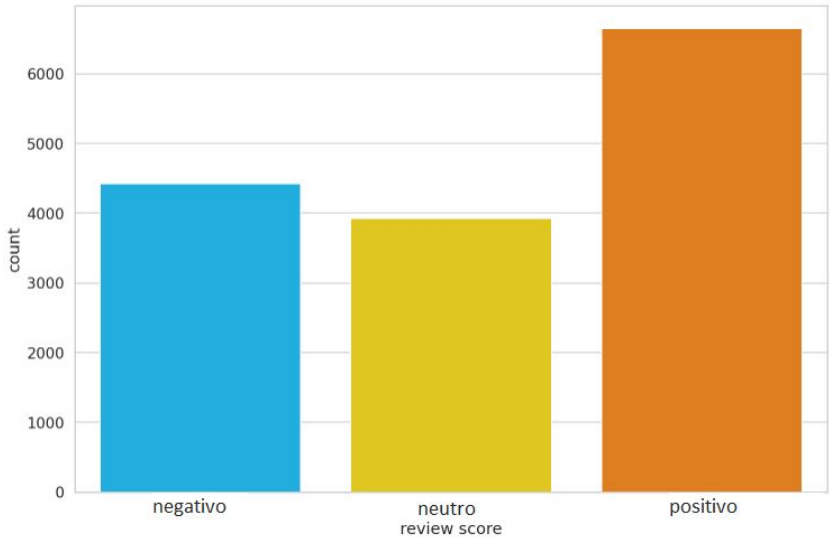
Métodos	**Medida F (%)**
Support vector machine	60,89
Naive bayes	59,12
Logistic regression	**64,87**
Multilayer perceptron	64,60
Decision tree	54,50
Random forest	62,18

Tabela 3: Resultados obtidos pelo autor do corpora TweetSentBr (BRUM, 2017).

A Tabela 3 contém os resultados da métrica Medida F obtidos pelo autor do corpora TweetSentBr (BRUM, 2017) e descrito em seu artigo. Com isso, é possível afirmar que o BERT é proeminente em relação aos principais algoritmos de AM presentes na literatura. Contudo, ainda precisamos refinar e expandir nossos experimentos, para que sejam o mais preciso possível. Um exemplo de refinamento necessário é o balanceamento das classes do corpora, que como

mostra a Figura 1, estão desbalanceados, o que pode deixar o modelo tendencioso a classe que possui mais entradas.

Figura 1: Quantidades de tweets por classe do corpora TweetSentBr.



4. CONCLUSÕES

Com base nos estudos e experimentos realizados, podemos afirmar que o modelo BERT-base para múltiplos idiomas se mostrou bastante promissor a alcançar bons resultados na tarefa de AS no nível de documento. Como trabalhos futuros pretendemos utilizar o modelo pré-treinado em português para reconhecimento de entidades nomeadas desenvolvido por (SOUZA, 2019) com um *fine-tuning* para AS, procurar por diferentes corpora e balanceá-los antes de testar o modelo pré-treinado para o português, para que os experimentos se tornem mais precisos e confiáveis.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BRUM, Henrico Bertini; NUNES, Maria das Graças Volpe. Building a sentiment corpus of tweets in brazilian portuguese. **arXiv preprint arXiv:1712.08917**, 2017.

CHOWDHURY, G. Natural language processing. **Annual Review of Information Science and Technology**, 37. pp. 51-89, 2003.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

GONÇALVES, P., ARAÚJO, M., BENEVENUTO, F., & CHA, M. Comparing and combining sentiment analysis methods. **COSN**, 2013.

LI, Susan. **Multi Class Text Classification With Deep Learning Using BERT**. Medium, 02 ago. 2020. Acessado em 25 set. 2020. Online. Disponível em:

https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. Portuguese named entity recognition using BERT-CRF. **arXiv preprint arXiv:1909.10649**, 2019.