

Evento: XXVIII Seminário de Iniciação Científica
ODS: 9 - Indústria, Inovação e Infra-estrutura

IMPLEMENTAÇÃO DE REDE NEURAL ARTIFICIAL APLICADA NA ANÁLISE DE SENTIMENTOS EM TEXTOS¹

IMPLEMENTATION OF AN ARTIFICIAL NEURAL NETWORK APPLIED IN THE ANALYSIS OF FEELINGS IN TEXTS

Cristiano Alex Künas², Leandro Perius Heck³, Edson Luiz Padoin⁴

¹ Trabalho de conclusão desenvolvido no curso de Ciência da Computação da UNIJUÍ

² Aluno do Curso de Graduação em Ciência da computação da UNIJUÍ, cristiano.kunas@sou.unijui.edu.br.

³ Aluno do Curso de Graduação em Ciência da computação da UNIJUÍ, leandro.h@sou.unijui.edu.br.

⁴ Professor Doutor do Departamento de Ciências Exatas e Engenharias da UNIJUÍ, Orientador, padoin@unijui.edu.br

1. INTRODUÇÃO

Diariamente, inúmeras informações circulam pela internet através de sites, blogs e redes sociais. O conteúdo presente vai desde análises e comentários de filmes até conversas e experiências dos seus usuários. Nesse sentido, usando os recursos tecnológicos podemos armazenar, recuperar e analisar uma enorme quantidade destes dados eficientemente. Uma questão é que esses dados não encontram-se estruturados de forma compreensível para um único sistema computacional.

As opiniões são tão importantes para quase todas as atividades humanas que, sempre que precisamos tomar uma decisão, queremos ouvir as opiniões de outros (LIU et al., 2010). Indivíduos e organizações estão cada vez mais usando o conteúdo dessas mídias para a tomada de decisões. Porém, os sistemas computacionais ainda encontram grande dificuldade em compreender como os clientes se sentem em relação aos produtos e serviços de uma certa companhia e em algumas situações, apenas alguns pontos do texto são relevantes: sobre quem se fala, e se o que se fala é bom ou ruim (SANTOS, 2017). A Inteligência Artificial (IA) pode ser aplicada para facilitar a compreensão desses dados, de tal maneira que possamos usar uma sentença em linguagem natural como entrada e extrair um conjunto de dados na saída. Desta forma, a utilização de uma Rede Neural Artificial (RNA) pode auxiliar na extração automática do sentimento ou sensação de sentenças (BRUM, 2015).

Dentre as diferentes subáreas de estudo que o campo de AS apresenta, a tarefa mais presente na literatura e que abordaremos neste trabalho é a análise de polaridade de um documento, que objetiva classificar textos em uma escala entre positivo e negativo, uma vez que essa área de pesquisa é de grande relevância para consumidores e organizações. Assim sendo, o presente trabalho almeja i) criar uma rede neural recorrente do tipo *Long Short-Term Memory*; ii) realizar o seu treinamento utilizando uma base de dados pública; e iii) aplicá-la no reconhecimento de sentimentos expressados em textos e/ou postagens.

O restante do trabalho está organizado da seguinte forma. A Seção 2 apresenta a metodologia utilizada na implementação e o ambiente de execução utilizado na realização dos testes. Resultados são discutidos na Seção 3, seguidos das considerações finais e perspectivas de trabalhos futuros.

2. METODOLOGIA

Evento: XXVIII Seminário de Iniciação Científica

ODS: 9 - Indústria, Inovação e Infra-estrutura

A implementação da RN deste trabalho é feita usando a linguagem de programação *Python*. Os principais módulos utilizados são: *Keras*, *Tensorflow* e *Scikit-Learn*. O *Keras* é usado para modelar a RN testada neste trabalho. Este é uma API de alto nível, desenvolvida com foco no usuário, para permitir experimentação rápida. Suporta diversas configurações de redes e funciona tanto com CPU quanto com GPU. O *Tensorflow* é o *back-end* padrão do *Keras*. Este é uma biblioteca de software de código aberto para computação numérica usando gráficos de fluxo de dados. É flexível e permite um rápido e fácil interfaceamento com CUDA (*Compute Unified Device Architecture*). Para dividir os dados usados neste trabalho, em treino e teste, usa-se o *Scikit-Learn*. Este é um módulo de aprendizado de máquinas, é um ecossistema *Python* para computação científica criado sobre o *SciPy*.

A base de dados usada nos experimentos deste trabalho foram elaborados no trabalho de Maas *et al.* (2011). Esta base possui uma coleção de 50.000 avaliações do IMDb. Este conjunto de dados é formado por sentenças escritas no idioma inglês. As críticas estão bem balanceadas em 2 classes possíveis: positiva (*pos*) e negativa (*neg*). O pré-processamento destes dados é feito em etapas com objetivo de remover ruído presente nas sentenças e codificar os textos de maneira que seja possível usá-los no treinamento de redes LSTMs. São elas: i) caracteres especiais e sinais de pontuação são removidos; ii) todas as palavras são escritas em letras minúsculas; iii) as *stopwords* são removidas; iv) as sequências são limitadas a um tamanho fixo (300 palavras); v) os dados são divididos em treino e teste.

A criação do modelo de RN é definido como uma sequência de camadas. A classe *Model* da API funcional da biblioteca *Keras* é utilizada no desenvolvimento do modelo de dados abordado neste trabalho. A compilação do modelo configura o processo de aprendizado. Ele define o otimizador (*adam*), a função de perda (*binary_crossentropy*) e as métricas (*accuracy*). Para treinar o modelo, os dados são divididos em lotes (*batch_size*), iterando repetidamente em todo o conjunto de dados para um determinado número de épocas. O modelo é avaliado por meio da função *evaluate()*. Essa função gerará uma predição para cada entrada e saída e coletará pontuação, realiza uma média sobre o valor de *perda* e *acurácia*.

O ambiente de execução é composto de um equipamento com um processador *Intel Core i7-9750* com 6 *cores* (12 *threads*) de 2.60 GHz de frequência. Este equipamento possui 16GB de Memória RAM DDR4, GPU NVIDIA GeForce GTX RTx 2060 com 6GB de GDDR6 e 1920 CUDA *cores*, utilizou-se do sistema operacional Linux Ubuntu 18.04.3 LTS com versão de *kernel* 5.0.0-37. A versão do NVIDIA CUDA *Compiler* utilizada foi a 10.0.130.

3. RESULTADOS E DISCUSSÃO

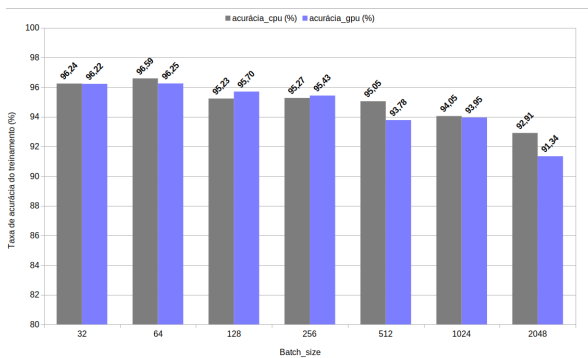
A Figura 1(a) apresenta a *acurácia* (eixo *y*) alcançada a partir do dataset de treinamento (40000 registros), com diferentes tamanhos de *batch*, nas execuções tanto em CPU quanto em GPU. A média de perda de cada modelo é demonstrada na Figura 1(b). Percebe-se que conforme se aumenta o *batch_size*, é diminuída a acurácia para o modelo, ou seja, conforme o número de instâncias para cada atualização de gradiente aumenta, a precisão do modelo é reduzida, e conseqüentemente a taxa de perda aumenta.

A Figura 2(a) apresenta a *acurácia* (eixo *y*) obtida a partir do *dataset* de validação (10000 registros), novamente variando o *batch_size* (32~2048) e executando tanto em CPU quanto em GPU. Esta métrica pode ser considerada a mais importante, visto que é a partir dela que definimos a precisão

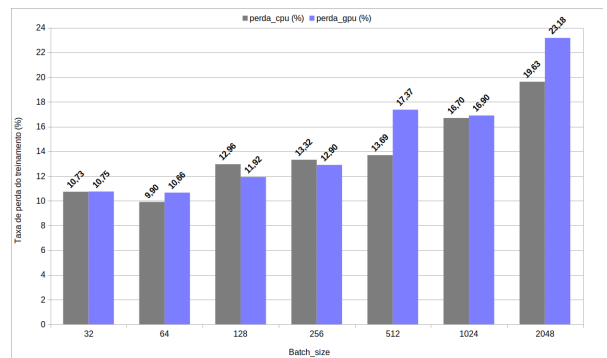
Evento: XXVIII Seminário de Iniciação Científica
 ODS: 9 - Indústria, Inovação e Infra-estrutura

do nosso modelo. A *acurácia* de validação nos informa o percentual que nosso modelo acertou na predição em relação ao rótulo, isto é, a partir de cada novo registro do *dataset*, para quantos o modelo foi capaz de definir o rótulo correto. Em nossos testes, para os diferentes tamanhos de *batch*, nota-se que a *acurácia* se mantém estável, variando entre 87~89%. Já a média de *perda*, Figura 2(b), apresenta um decréscimo gradual conforme aumenta o tamanho do *batch*.

Figura 1: Métricas dos dados de treinamento das execuções em CPU e GPU



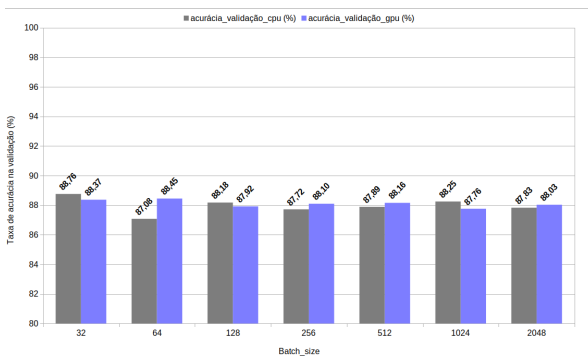
(a) *Acurácia* obtida



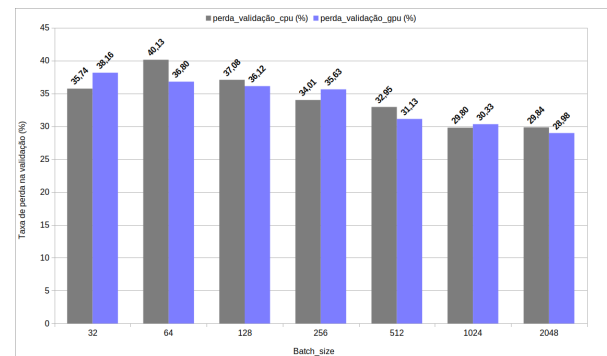
(b) *Perda* observada

Fonte: Próprio Autor.

Figura 2: Métricas dos dados de validação das execuções em CPU e GPU



(a) *Acurácia* obtida



(b) *Perda* observada

Fonte: Próprio Autor.

4. CONSIDERAÇÕES FINAIS

Este trabalho abordou o uso de redes neurais recorrentes *Long Short-Term Memory* na classificação de sentimentos em sentenças. Utilizou-se diferentes configurações para analisar o desempenho da aplicação executada em arquiteturas CPU e GPU. No que tange a taxa de *acurácia*, para dados de treinamento, em ambas arquiteturas, a melhor taxa é observada para *batch* de tamanho 64. Nos dados

Evento: XXVIII Seminário de Iniciação Científica

ODS: 9 - Indústria, Inovação e Infra-estrutura

de validação, em arquitetura CPU, a melhor taxa é registrada para *batch* de tamanho 32. Já para arquitetura GPU, a melhor taxa é obtida para *batch* de tamanho 64. A proposta apresentou bom desempenho no que se refere à precisão da RNA, esta métrica ainda pode ser melhorada através de novas modificações em seus parâmetros e novos testes.

Como trabalhos futuros, o primeiro a ser feito é usar *word embeddings* pré-treinados. Além disso, é possível estudar a influência de outros hiperparâmetros, como taxa de aprendizagem, taxa de *Dropout* e Função de Ativação. Também é possível avaliar base de dados maiores e em outros idiomas. Uma outra análise que pode ser feita é adicionar outras classes de classificação além de positivo e negativo, como por exemplo, muito positivo, muito negativo e neutro. Como também avaliar outras fontes de dados como *twitter*. Outra abordagem a ser estudada é uma melhor separação das bases de treino, teste e validação.

REFERÊNCIAS BIBLIOGRÁFICAS

BRUM, Henrico Bertini. **Análise de sentimentos para o português usando redes neurais recursivas**. 2015. *Monografia* (Bacharel em Ciência da Computação), Universidade Federal do Pampa, Alegrete, Brazil.

DOS SANTOS, Igor Pedro Pinto. **Análise de sentimento usando redes neurais de convolução**. 2017. *Tese de Doutorado*. Dissertação (Mestrado em Engenharia Eletrônica), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil.

LIU, Bing et al. **Sentiment analysis and subjectivity**. *Handbook of natural language processing*, v. 2, n. 2010, p. 627-666, 2010.

MAAS, Andrew et al. **Learning word vectors for sentiment analysis**. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011. p. 142-150.

Parecer CEUA: 2208566