UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIENTÍFICA E TECNOLÓGICA - 2019

ESTUDOS DE TÉCNICAS PARA PROCESSAMENTO DE LINGUAGEM NATURAL

Sarah Pereira Cerqueira[1] e Angelo Amâncio Duarte[2]

1. Bolsista PIBIC/CNPq, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: sarahecomp@gmail.com

2. Orientador, Departamento de Tecnologia, Universidade Estadual de Feira de Santana, e-mail: angeloduarte@uefs.br

PALAVRAS-CHAVE: Inteligência Artificial; Processamento de Linguagem Natural; Deep Learning.

INTRODUÇÃO

Processamento de Linguagem Natural (do inglês *Natural Language Processing* - NLP) é uma área da inteligência artificial que tem se ocupado com o desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em língua natural.

Linguagem Natural, seja ela falada ou escrita é o meio mais natural de comunicação entre humanos, e o modo mais escolhido para produzir documentos. Uma vez que os computadores têm um importante papel na preparação, aquisição, transmissão, monitoramento, armazenamento, análise e transformação de informação, capacita-los a entender e gerar informação expressa em linguagem natural é cada vez mais necessário (WEISCHEDEL, 2003).

Entretanto, desenvolver aplicações que realizam tarefas NLP envolve uma série de dificuldades associadas à linguagem como a diversidade de idiomas e dialetos, gírias, erros gramaticais, ambiguidade, abreviações, particularidades regionais, etc. Além disso, há pouca disponibilidade de ferramentas para NLP em português.

No Laboratório de Computação de Alto Desempenho (LACAD) da Universidade Estadual de Feira de Santana (UEFS), o projeto em andamento na área de Patologia Renal necessita da análise de laudos médicos obtidos por patologistas e clínicos. Assim, o objetivo deste trabalho é o estudo de algoritmos, técnicas e ferramentas de NLP que possam escalar a análise de texto, suprindo as demandas dos projetos do LACAD, e agregando a tecnologia ao laboratório.

METODOLOGIA

O trabalho usou da infraestrutura do Laboratório de Computação de Alto Desempenho (LACAD), localizado na sala I12 do LABOTEC 3, e foi desenvolvido com as seguintes etapas principais:

1. Revisão bibliográfica
    a. Aprendizagem dos conceitos de NLP;
    b. Estudo das plataformas para processamento de linguagem natural atuais;

2. Seleção da plataforma de NLP para o LACAD
    a. Definição das métricas que serão usadas para comparar as plataformas encontradas;
    b. Experimentos e comparação de desempenho das plataformas;
3. Divulgação dos conhecimentos
    a. Produzir e publicar no site do LACAD tutoriais de uso das ferramentas encontradas;
    b. Treinar pesquisadores do LACAD que necessitem de NLP em seus projetos;
4. Divulgação dos resultados através de palestras ou publicação de artigos.

RESULTADOS E DISCUSSÃO

Utilizando de artigos científicos de domínio público, foi feito um estudo e análise de técnicas que são utilizadas para processar textos em linguagem natural. Como resultado descobriu-se que em geral, o NLP em textos são feitos basicamente em duas fases principais, a primeira sendo a representação de palavras e a segunda treinamento de uma rede neural.

A representação de palavras é necessária para possibilitar que os textos sirvam de entrada para as redes neurais. Em geral, verificou-se ser útil representar as palavras como vetores, por eles possuírem uma interpretação atraente e intuitiva, podendo ser objeto de operações como adição, subtração, mensurar distância, etc, e prestam-se bem para serem usados em algoritmos e estratégias de aprendizagem profunda (ALMEIDA, XEXÉO, 2019).

Atualmente o paradigma mais utilizado para representar palavras é chamado de *word embeddings* que se baseia na hipótese distribucional. Tal hipótese, assume que palavras com significado semelhante tendem a ocorrer em contextos similares (ALMEIDA, XEXÉO, 2019; YOUNG *et al*. 2017).

O modelo de *word embeddings* mais citado na literatura é o *Word2Vector* (Mikolov *et al*. em 2013), seguido pelo *Global Vectors* (GloVe) (PENNINGTON *et. al* 2014). Ambos modelos conseguem mapear as palavras para um espaço significativo, onde a distância entre às palavras está relacionada à sua semelhança semântica. A diferença entre ambos modelos se dar quanto a estratégia utilizada em sua construção, enquanto o *Word2Vector* é um modelo baseado em predição (utiliza de algoritmos de treinamento), o GLoVe é baseando em contagem (conta e faz uma estatística de quantas vezes uma palavra co-ocorre com suas palavras vizinhas).

Na fase de treinamento de uma rede neural, é comum a utilização das redes neurais convolucionais (do inglês *convolutional neural network* CNNs) e as redes neurais recorrentes (do inglês *recurrent neural network* RNNs), ambas redes têm produzido o estado da arte em tarefas de NLP.

É uma tendência que pesquisadores escolham CNN para tarefas de classificação, e RNN para tarefas de modelação sequencial devido à natureza das arquiteturas. No estudo feito por Yin *et al*. (2017), que faz uma comparação entre CNN e RNN no contexto de NLP, é chegada a conclusão que a melhor arquitetura para uma determinada tarefa, vai depender do quão importante é a compreensão semântica de toda sentença.

Apesar dos inúmeros desafios relacionados ao processamento de linguagem natural, existem muitas ferramentas que dão suporte ao desenvolvimento de tarefas de NLP. Entretanto, a maioria das ferramentas disponíveis dão suporte maior para o inglês, em detrimento de outros idiomas.

Com a finalidade de adequar os equipamentos do LACAD à nova área de atuação instalou-se as ferramentas NLTK e Spacy. A ferramenta NLTK foi escolhida por ser a

mais popular e completa para fazer tarefas de NLP, e a ferramenta Spacy foi escolhida devido ao seu desempenho, o qual a está fazendo ficar cada vez mais popular também. Ambas ferramentas foram instaladas e testadas.

A fim de complementar esse estudo, implementou-se um modelo capaz de fazer análise de sentimento em comentários. O conjunto de dados usado foi do Yelp, que é uma empresa que ajuda usuários a encontrar serviços e estabelecimentos. Devido ao programa *Yelp Dataset Challenge*, o Yelp disponibilizou no site Kaggle, subconjuntos de dados para que estudantes fizessem pesquisa e análise de seus dados, compartilhando posteriormente suas descobertas. Para essa pesquisa o conjunto de dados escolhido foi o de avaliações de diversos serviços, com mais de quinhentos mil comentários. Os comentários foram classificados em positivo, negativo e neutro a partir da quantidade de estrelas dadas pelos usuários que fizeram as avaliações.

Primeiramente os dados foram pré-processados retirando caracteres especiais, em seguida transformou-se cada comentário em vetores de token, onde cada palavra das sentenças representava um token, e para cada palavra do vocabulário foi criado um índice único. Por fim, antes de serem treinados os vetores de sentença foram ajustados para terem o mesmo tamanho.

A arquitetura do modelo possui quatro camadas, sendo que a primeira camada é responsável por criar uma representação distribuída para cada index de entrada do modelo, sua dimensionalidade de saída é de 128. A segunda camada LSTM é responsável por aprender a classificar as sentenças, sua dimensionalidade de saída foi configurada para 300. A terceira camada é a de regularização que utiliza uma técnica chamada *dropout*, cuja finalidade é reduzir o *overfitting* do modelo. Já a última camada, força que a saída seja a probabilidade dos dados de pertencer a uma das classes: positivo, negativo e neutro.

Na configuração de treinamento do modelo outros parâmetros foram definidos conforme a Tabela 1.

Tabela 1. Configuração de treinamento do modelo LSTM

Função de Perda	Categorical Cross Entropy
Otimização	Adam
Métrica	Acurácia
Batchs	32
Epoch	3

Para o treinamento do modelo somente oitenta mil comentários e as 5000 palavras mais frequentes foram considerados. Dos oitenta mil dados 20% foram separados para teste, e como resultado conseguiu-se uma acurácia de 84.53% nos dados de treinamento e 84.08% nos dados de testes. O que são resultados bons, considerando a quantidade de dados utilizada, certamente se a quantidade de dados fosse aumentada os resultados também melhorariam, entretanto demandaria muito mais tempo de treinamento, que nesse experimento consumiu 2.62 horas em um processador i5 com 8GB de RAM.

Aumentar o número de epochs também ajudariam no melhoramento do modelo, entretanto é necessário ter cuidado com overfitting, uma dica é observar a acurácia do

teste, se ela for muito menor que a acurácia dos dados de treinamento, certamente houve um overfitting.
O código implementado está devidamente comentado em português, e é um bom exemplo para quem necessita aprender a implementar uma rede neural com Keras. No final do código ainda é possível testar o modelo com comentários que não fizeram parte do treinamento, verificando se o comentário é positivo, negativo ou neutro.

CONCLUSÃO

Visando agregar a tecnologia de processamento de linguagem natural ao LACAD, a fim de suprir as demandas do laboratório, realizou-se um estudo das ferramentas e técnicas que mais estão sendo utilizadas para processar textos. Como resultado tem-se um relatório com um resumo dos métodos mais utilizados para processar texto, e as ferramentas mais populares utilizadas para desenvolver aplicações com esta tecnologia. Além disso foi possível desenvolver algo prático, aplicando o conhecimento obtido para fazer análise de sentimento em comentários.
A disponibilidade de estudos e recursos de NLP voltados ao português ainda é muito escasso. O que dificulta a construção de aplicações que necessitam processar conteúdos expressos em português.
Em relação à escolha de qual técnica utilizar, está sempre dependerá do tipo de tarefa de processamento de linguagem que será implementada, e de encontrar a configuração correta para aquela tarefa.

REFERÊNCIAS

ALMEIDA, F. XEXÉO, G. 2019. *Word Embeddings: A Survey*. Rio de Janeiro: Universidade Federal do Rio de Janeiro. [ONLINE] Disponível em: <https://arxiv.org/pdf/1901.09069.pdf> Acessado 11 de fevereiro de 2019.

MIKOLOV, T. et al. 2013. *Efficient Estimation of Word Representations in Vector Space.* Mountain View: Google Inc. [ONLINE] Disponível em: <https://arxiv.org/pdf/1301.3781.pdf> Acessado 8 de março de 2019.

PENNINGTON, J., SOCHER, R., MANNING, C. D. 2014. *GloVe: Global Vectors for Word Representation*. Stanford: Stanford University [ONLINE] Disponível em: <https://nlp.stanford.edu/pubs/glove.pdf > Acessado 19 de março de 2019.

WEISCHEDEL, R. et al. 2013. *White Paper on Natural Language Processing.* [ONLINE] Disponível em: <https://www.researchgate.net/publication/234826005>. Acessado 14 de setembro de 2018.

YIN, W. et al. 2017. *Comparative Study of CNN and RNN for Natural Language Processing.* CoRR: abs / 1702.01923. [ONLINE] Disponível em: <https://arxiv.org/pdf/1702.01923.pdf> Acessado 17 de dezembro de 2018.

YOUNG, T. et al. 2017. *Recent Trends in Deep Learning Based Natural Language Processing*. CoRR: abs/1708.02709. [ONLINE] Disponível em: <https://dblp.org/rec/bib/journals/corr/abs-1708-02709> Acessado 14 de setembro de 2018.