CONSTRUÇÃO DE UM *CORPUS* LINGUÍSTICO PARA ANÁLISE DE SENTIMENTOS EM RESENHAS DE SUPERMERCADOS E TESTES USANDO APRENDIZADO PROFUNDO[1]

Matias Giuliano Gutierrez Benitez[2], Rui Jorge Tramontin Júnior[3].

[1] Vinculado ao projeto "Estudo Comparativo de Técnicas para Análise de Sentimento e Desenvolvimento de uma Biblioteca de Programação para Análise de Textos em Língua Portuguesa"
[2] Acadêmico (a) do Curso de Ciência da Computação – CCT – Voluntário PROVIC
[3] Orientador, Departamento de Ciência da Computação – CCT – rui.tramontin@udesc.br

O Processamento de Linguagem Natural (PLN) é uma vertente da Inteligência Artificial que ajuda computadores a entender, interpretar e manipular a linguagem humana. O PLN resulta de diversas disciplinas, incluindo Ciência da Computação e Linguística Computacional, que buscam preencher a lacuna entre a comunicação humana e o entendimento dos computadores. Dentro do PLN existe uma subárea chamada Análise de Sentimentos (AS) em ela se estudam as opiniões, atitudes e emoções das pessoas em relação a uma entidade utilizando o poder computacional. Esta entidade pode ser representada por indivíduos, eventos ou tópicos.

O objetivo da AS é encontrar opiniões, identificar o sentimento que estas opiniões expressam e, por último, classificar sua polaridade. Existem três níveis de classificação, o primeiro é em *nível de sentença*, que visa classificar o sentimento expresso em uma frase. O segundo é em *nível de documento*, que tem como objetivo classificar um conjunto de frases e dar para elas uma polaridade positiva ou negativa. Em último lugar, há o *nível de aspecto*, que utiliza o conceito de entidades para fazer a classificação. Esse nível identifica uma entidade dentro do texto e procura por opiniões referentes a tal entidade.

Há diversas aplicações possíveis para a AS, como por exemplo, análise de comentários na rede social *Twitter* para predizer o resultado das eleições, identificar sentimentos em músicas para aplicá-los em sistemas de recomendação e análise da opinião pública sobre um determinado produto para o desenvolvimento de estratégias de *marketing*. Em resumo, onde existe uma opinião dirigida a algum tema em especifico, é possível aplicar a AS. O foco deste trabalho é na análise de resenhas de supermercados.

Técnicas de AS podem ser divididas em aprendizado de máquina, abordagens baseadas em léxicos de sentimentos, técnicas baseadas em ontologias e aprendizado profundo. Muitas dessas técnicas utilizam um *corpus*: uma coleção de textos em formato eletrônico que representam um idioma. Embora a pesquisa em AS no idioma inglês seja extensa, especialmente na construção de *corpora*, a língua portuguesa ainda carece de contribuições nessa área. Portanto, um dos objetivos deste trabalho é o desenvolvimento de um *corpus* extraído de resenhas de supermercados.

Para a construção do *corpus*, os textos foram coletados do *Google Places*, visto que é uma fonte com muitas resenhas de produtos e serviços. A escolha do domínio "supermercados" se justifica pelo fato de ser um tipo de estabelecimento que existe em qualquer cidade. Depois da coleta dos dados, uma etapa de revisão foi executada e a versão final do *corpus* foi compilada. O *corpus* possui um total de 7121 textos, onde 1082 tem polaridade negativa, 1004 são neutros e 5035 são positivos.

Para analisar a utilidade do *corpus*, o mesmo foi utilizado como base de treinamento e testes na abordagem chamada de *aprendizado profundo*. Tal abordagem aplica redes neurais artificiais de modo que possam aprender utilizando múltiplas camadas inspiradas em um cérebro biológico. A rede é composta por várias unidades de processamento, organizadas em camadas, chamadas neurônios. O aprendizado ocorre através do ajuste de pesos entre os neurônios, reestruturando o processo de aprendizado, tal como fazem os cérebros biológicos. Neste trabalho, foi utilizado o *Perceptron Multicamadas*, um tipo de rede neural que pode aprender uma aproximação de uma função não linear a partir de um conjunto de treinamento.

Portanto, uma vez treinada com o *corpus* desenvolvido, a rede neural deve ser capaz de identificar a polaridade de uma dada resenha. Para este trabalho, considera-se valores discretos para representar as polaridades, ou seja, a polaridade de um documento é classificada como negativa (-1), neutra (0) ou positiva (1).

Para que o *corpus* pudesse ser usado nos testes, o mesmo foi submetido a uma etapa de *pré-processamento* visando normalizar o texto de todos os documentos: conversão para letras minúsculas, lematização das palavras e remoção de palavras que não agregam valor à polaridade (*stop-words*). Para o treinamento e teste, foram testadas três redes iguais, mas cada uma com funções de ativação diferentes: *Sigmóide*, *Tanh* (Tangente hiperbólica) e *ReLU* (*Rectified Linear Unit* ou Unidade Linear Retificada). Foi utilizado o método de validação cruzada em 10 passos (*10-fold cross-validation*), e as seguintes medidas foram coletadas: acurácia, precisão, medida F1 e *recall*. A implementação foi feita na linguagem Python usando o pacote *scikit-learn* (versão 1.0.2).

Os testes com as três configurações alcançaram resultados muito próximos entre si, conforme mostra a Tabela 1. Estes foram os resultados iniciais e ainda há espaço para melhorias, considerando que podem ser feitos ajustes nas etapas de pré-processamento do texto bem como nos parâmetros utilizados para a configuração da rede neural.

É preciso salientar que muito do esforço realizado neste trabalho foi concentrado na construção do *corpus*. A extração, seleção, revisão e anotação dos textos foram tarefas que consumiram muito tempo. Por fim, consideramos os resultados obtidos uma importante contribuição para a área, sobretudo devido à carência de *corpora* aplicados à AS na língua portuguesa. Todos os códigos e o *corpus* desenvolvido estão disponíveis publicamente no GitHub.

Tabela 1. *Resultados dos testes.*

	Sigmóide	***Tanh***	**ReLU**
Acurácia	0,764	0,764	**0,784**
Precisão	0,768	0,765	**0,785**
Recall	0,764	0,764	**0,784**
F1	0,765	0,764	**0,784**

Palavras-chave: Análise de Sentimentos. *Corpus*. Aprendizado Profundo.