ANÁLISE DE SENTIMENTOS EM RESENHAS DE SUPERMERCADOS: CONSTRUÇÃO DE UM *CORPUS* LINGUÍSTICO PARA O PORTUGUÊS E TESTES USANDO APRENDIZADO DE MÁQUINA[1]

Vinicius Takeo Friedrich Kuwaki[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

Nos últimos anos, com a popularização das redes sociais, o número de informações textuais disponíveis para a extração de conhecimento estruturado cresceu em escalas ilimitadas. A análise de sentimentos (AS) é uma subárea do processamento de linguagem natural que busca determinar se uma informação textual expressa uma opinião com relação a uma entidade. Uma opinião é definida como sendo um sentimento ou avaliação expressa por uma entidade com relação a outra entidade ou evento. Uma sentença que expressa uma opinião é dita subjetiva, enquanto que sua contraparte é chamada de objetiva. As técnicas na AS buscam: (1) determinar se uma informação textual é subjetiva ou objetiva, e; (2) caso seja subjetiva, determinar sua polaridade, i.e., um valor contínuo ou discreto que represente o grau de positividade ou negatividade. A AS pode ser realizada em três níveis: (1) em nível de documento, onde um conjunto de sentenças é analisado; (2) em nível de sentença, onde um conjunto de palavras é analisado, ou; (3) em nível de aspecto, onde uma característica de uma entidade é colocada sob análise em um documento ou sentença.

São várias as aplicações possíveis para a AS: manipulação de ativos no mercado de ações, análise da opinião pública com relação a um evento, tal como a pandemia de COVID-19 ou as eleições, por exemplo, além de outras como a análise de comentários de vídeos no YouTube e até mesmo resenhas de filmes, músicas e produtos em geral. O foco deste trabalho é na análise de resenhas de supermercados.

Técnicas de AS podem ser divididas em aprendizado de máquina, abordagens híbridas, abordagens baseadas em léxico, 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. Os textos foram extraídos através de técnicas de *web scraping*, que consiste na prática de coletar dados através de outros meios que não sejam uma interface de programação. 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.

O *corpus* foi então testado por três diferentes abordagens de aprendizado de máquina: Regressão Logística, *Naive Bayes* e *Support Vector Machines* (SVM). Tais abordagens são algoritmos ditos *supervisionados*, ou seja, necessitam de dados anotados (tal como um *corpus*) para

que sejam treinados e posteriormente possam fazer suas previsões. Em linhas gerais, esse tipo de algoritmo visa computar a probabilidade de uma dada entidade de pertencer ou não a uma dada classe. Para este trabalho, considera-se valores discretos (-1, 0 e 1) para representar as classes. Os algoritmos são treinados para classificar um documento como pertencendo a uma dessas três classes. Portanto, a polaridade de um documento é classificada como negativa (-1), neutra (0) ou positiva (1).

Para testar os três algoritmos, o *corpus* foi submetido a uma etapa de *pré-processamento* visando converter os documentos a uma formal normal: conversão para letras minúsculas, lematização das palavras e remoção de palavras comuns (*stop-words*). Para o treinamento e teste dos algoritmos, 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).

Todos os métodos alcançaram resultados similares, conforme mostra a Tabela 1. O modelo *Naive Bayes* apresentou resultados levemente melhores que os demais métodos. Embora os resultados já sejam considerados bons, ainda há espaço para melhorias, considerando ajustes nas etapas de pré-processamento do texto e ajustes nos parâmetros dos algoritmos utilizados.

É importante destacar que muito do esforço realizado foi concentrado na seleção, revisão e anotação dos textos para construção do *corpus*. Consideramos os resultados obtidos uma importante contribuição para a comunidade acadêmica, sobretudo devido à carência de *corpora* para treinamento de sistemas de AS na língua portuguesa. Todos os códigos e o *corpus* desenvolvido estão disponíveis publicamente em um repositório do GitHub.

Tabela 1. *Resultados dos testes.*

	Regressão Logística	***Naive Bayes***	**SVM**
Acurácia	0,768	**0,825**	0,794
Precisão	0,796	**0,804**	0,769
Recall	0,768	**0,825**	0,794
F1	0,779	**0,808**	0,768

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