ANÁLISE DE SENTIMENTOS: INTERPRETANDO OPINIÕES NAS AVALIAÇÕES DE SUPERMERCADOS COM LÉXICOS DE SENTIMENTOS[1]

Mateus Nepomuceno Ladeira[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

Com o crescente desenvolvimento de novas tecnologias, as tarefas do dia a dia têm se tornado cada vez mais simples. Nesse contexto, aplicações *mobile* permitem que a sociedade tenha acesso aos mais diversos serviços *online*. Como consequência, ocorre um constante aumento do número de opiniões disponíveis na internet, tornando-se fundamental o desenvolvimento de aplicações que as analisem de forma automatizada. Nesse sentido, diversos estudos sobre Análise de Sentimentos (AS), ramo do processamento de linguagem natural, estão em desenvolvimento.

Esse campo de estudos busca verificar a ocorrência de opiniões em textos, e, quando houver, classificá-las de acordo com sua polaridade. Quando não há opinião no texto analisado, ele é classificado como objetivo; caso contrário, ele é subjetivo. Após essa primeira verificação, constatando-se a subjetividade, outra análise deve ser feita, verificando se a opinião expressa pelo conteúdo é positiva, negativa ou neutra.

Existem diferentes métodos para AS de um texto, dentre os quais, dois se destacam: classificação por meio de um recurso léxico e análise com aprendizado de máquina. Um *léxico de sentimentos* consiste em um conjunto de palavras que possuem uma polaridade atribuída, e é uma ferramenta muito utilizada para mineração de opinião. Alguns léxicos de sentimentos de propósito geral podem ser encontrados nos trabalhos correlatos, o que significa, portanto, que não são de domínio específico.

Considerando que palavras podem ter significados diferentes dependendo do domínio, este estudo visa desenvolver um léxico de sentimentos específico para a análise de resenhas de supermercados. A escolha do domínio "supermercados" se justifica pelo fato de ser um tipo de estabelecimento que existe em qualquer cidade. Além disso, os resultados aqui obtidos poderão futuramente ser mais facilmente adaptados a resenhas de outros tipos de produtos e serviços. O trabalho foi feito em três etapas: (1) construção de um *corpus*; (2) construção do léxico de sentimentos; (3) testes para análise do desempenho do léxico desenvolvido.

Um *corpus* é 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. Tais resenhas foram extraídas do *Google Places*, que é uma fonte que disponibiliza muitas avaliações de produtos e serviços. 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. Além de ser usado na construção do léxico polarizado, um *corpus* pode servir também como base para treinamento e testes de técnicas de aprendizado de máquina. Porém, esta tarefa está fora do escopo deste trabalho.

Após a construção do *corpus* específico ao domínio escolhido, o léxico de sentimentos foi desenvolvido. Como foi dito, um léxico de sentimentos é um dicionário ou lista de palavras com suas respectivas polaridades (positiva, negativa ou neutra). Todas essas palavras foram conferidas por dois anotadores que foram responsáveis por determinar sua polaridade. As polaridades são representadas por valores discretos (-1, 0 e 1), sendo -1 como negativa, 0 como neutra e 1 como positiva. No final, o léxico proposto teve um total de 1995 palavras anotadas, sendo 966 verbos, 784 adjetivos e 245 advérbios.

Para testar o léxico de sentimentos desenvolvido, foi feita uma comparação de seu desempenho em comparação com o LIWC-PT, um léxico de propósito geral. Para isso, foram extraídos todos os verbos, advérbios e adjetivos dos textos contidos no *corpus* e anotados manualmente de acordo com sua polaridade no domínio de "supermercados". Por exemplo, a palavra "espera" pode significar que o estabelecimento tem longas filas, o que indica uma caraterística negativa. Por outro lado, "espera" tem uma polaridade positiva conforme o LIWC-PT, supostamente por estar relacionado a "ter esperança".

Na abordagem proposta, o algoritmo conta o número de palavras positivas e negativas de cada sentença, invertendo sua polaridade quando se encontra um indicador de negação (tal como o "não"). Depois de calcular as polaridades de todas as sentenças, a soma de todos os valores determina a polaridade resultante do texto como um todo, ou seja, da resenha analisada.

Para realizar os testes, 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 sem sentido para a análise (*stop-words*). O algoritmo proposto foi testado aplicando-se o léxico proposto e o LIWC-PT. As seguintes medidas foram coletadas: acurácia, precisão, medida F1 e *recall*. A implementação foi feita na linguagem Python.

Como esperado, o algoritmo proposto utilizando o léxico específico obteve resultados melhores que o LIWC-PT, conforme mostra a Tabela 1. Isso confirma a evidência pontada pelos trabalhos correlatos, de que um recurso léxico específico a um domínio permite atingir resultados melhores. Embora os resultados já sejam considerados bons, ainda há espaço para melhorias, considerando ajustes nas etapas de pré-processamento do texto e ajustes no próprio algoritmo.

É importante destacar que muito do esforço realizado foi concentrado na seleção, revisão e anotação dos textos para construção do *corpus*, bem como na posterior construção do léxico polarizado. Consideramos os resultados obtidos uma importante contribuição para a comunidade acadêmica, sobretudo devido à carência de recursos léxicos para suporte a sistemas de Análise de Sentimento na língua portuguesa. Todos os códigos e recursos desenvolvidos estão disponíveis publicamente em um repositório do GitHub.

Tabela 1. *Resultados dos testes.*

	LIWC	**Léxico proposto**
Acurácia	0,560	**0,762**
Precisão	0,667	**0,774**
Recall	0,560	**0,762**
F1	0,594	**0,767**

Palavras-chave: Análise de Sentimentos. Abordagem Lexical. Léxico de Sentimentos.