Use of Natural Language Processing and Machine Learning in mining product reviews in an online store Uso do Processamento de Linguagem Natural e Aprendizado de Máquina na mineração de opiniões de produtos em loja virtual

Abstract: In this work a study and development of Natural Language Processing system using Supervised Machine Learning was carried out to promote the opinion mining in products' reviews in virtual stores. From the automatic retrieval of customer comments on the web, using a web crawler, the development followed with the pre-processing of the sentences and their classifications using the Logistic Regression algorithm. Was obtained as result a precision of 89% using 10-fold cross-validation.

Resumo: Neste trabalho realizou-se o estudo e o desenvolvimento de um sistema de Processamento de Linguagem Natural utilizando Aprendizado de Máquina Supervisionado para promover a mineração de opiniões em avaliações de produtos em lojas virtuais. A partir da obtenção automática dos comentários dos clientes na *web*, a partir de um robô, o desenvolvimento seguiu com o pré-processamento das sentenças e suas classificações usando o algoritmo de Regressão Logística. Obteve-se como resultado uma precisão de 89% utilizando validação cruzada.

Palavras-chave: Processamento Linguagem Natural, *Natural Language Processing*, Aprendizado de Máquina, *Machine Learning*, Mineração de Opiniões, *Opinion Mining*.

1. Introdução

Num mundo globalizado, as pessoas usam principalmente a comunicação por texto nas mídias sociais disponíveis na *web* como, por exemplo, fóruns, blogs e redes sociais, para compartilhar suas observações, fazer perguntas e se envolverem com outras pessoas em diálogos significativos sobre seus dilemas. Este compartilhamento gera um enorme volume de dados opinativos que são de grande interesse acadêmico, comercial e político por permitir, por exemplo, identificar tendências da opinião pública para fins de análise comportamental ou marketing. Essa análise é um dos objetivos da área de pesquisa intitulada Mineração de Opiniões, também conhecida como Análise de Sentimentos, e consiste numa subárea de pesquisa do Processamento de Linguagem Natural (PLN). A mineração de Opinião, de acordo com Medhat *et al*. (2014), permite identificar a polaridade de uma determinada sentença, ou seja, se ele está relatando algo positivo ou negativo.

Segundo Liu (2010), um dos principais desafios para a área é, certamente, o reconhecimento e a classificação automática de opiniões e fatos naturalmente escritos em língua natural. Portanto, pelo exposto, objetivou-se neste trabalho realizar uma estudo sobre o Processamento de Linguagem Natural e os algoritmos de Aprendizado de Máquina e como estes podem ser aplicados na mineração de opiniões de comentários de avaliações de produtos em lojas virtuais.

De acordo com Horrigan (2008), está se tornando evidente que as opiniões expressas na *web* podem influenciar leitores na formação de suas opiniões sobre algum tema, principalmente no que concerne a compra de um produto ou a contratação de um serviço.

Segundo Kozinets (2001), a etnografia na Internet, ou também chamada de netnografia, é uma metodologia qualitativa de pesquisa do consumidor que utiliza as informações disponíveis publicamente em fóruns *online* para identificar e compreender as necessidades e influências

decisórias de grupos de consumidores. Em comparação com a etnografia orientada para o mercado, a netnografia é muito menos demorada e capaz de ser conduzida de uma maneira totalmente discreta, ou seja, em comparação com grupos focais e entrevistas pessoais, a netnografia é muito menos intrusiva, pois é conduzida usando observações de consumidores em um contexto que não é fabricado pelo pesquisador de marketing.

Portanto, pelo exposto, justifica-se o desenvolvimento deste projeto pela sua temática, a qual vem ao encontro das expectativas do mercado em relação ao uso da Inteligência Artificial nos negócios, principalmente no que concerne ao tratamento de dados não estruturados para auxiliar a mineração de opiniões de produtos.

2. Metodologia

Neste trabalho foi realizada uma pesquisa descritiva (ANDRADE, 2002), na qual realizou-se a análise de dados utilizando os materiais e métodos descritos nesta seção.

Num primeiro momento foi necessário montar o conjunto de dados (*dataset*) com os comentários sobre um produto específico numa loja virtual, o qual foi realizado a partir da implementação e uso de um *Web Crawler* capaz de ler uma página *web* e identificar os dados necessários para a execução desse projeto, quais sejam: os comentários e a quantidade de estrelas atribuídas por cada usuário sobre o produto em questão. Nesse momento, foi obtido um total de 789 comentários, sendo 51 classificados como negativos (menos de 3 estrelas) e 738 classificados como positivos (3 ou mais estrelas). A fim normalizar os dados para um treinamento mais preciso do sistema de classificação, utilizamos 102 comentários, sendo metade com comentários positivos e metade com comentários negativos. Os dados foram salvos no formato CSV e trabalhados a partir da biblioteca Pandas.

Uma vez obtido o *dataset*, aplicou-se as técnicas de Processamento de Linguagem Natural relacionadas ao pré-processamento dos dados a fim de eliminar os acentos e pontuações das sentenças, assim como os termos que não possuíam qualquer significado semântico para a classificação. Em seguida, as sentenças pré-processadas foram transformadas em um formato de vetor para uso no classificador.

Por fim, foi realizado o treinamento e teste do *dataset*, usando o classificador binário de Regressão Logística, escolhido por apresentar melhores resultados quando comparado aos algoritmos de Aprendizado de Máquina Árvore de Decisão e Naive Bayes.

Para o desenvolvimento desse projeto utilizou-se a linguagem de programação Python (versão 3.7) e as bibliotecas NLTK (Natural Language Toolkit) para o processamento de linguagem Natural e Scikit-Learn para o Aprendizado de Máquina. Nas próximas seções são apresentadas as etapas realizadas.

2.1 Criação do *corpus* utilizando *Web Crawler*

Um *Web Crawler* (em português, rastreador da rede ou robô), é um programa de computador que navega pela rede mundial de forma automatizada para realizar a indexação de documentos, sendo muito utilizado pelos motores de busca atuais. O *Web Crawler* foi desenvolvido neste trabalhoo para criar um conjunto de dados com comentários de um produto específico em um site de e-commerce, em específico, uma impressora multifuncional que continha, na data da execução do *Web Crawler*, 789 avaliações.

A biblioteca utilizada para desenvolver o *Web Crawler* foi a Requests, bastante utilizada nessa área por permitir acessar conteúdos na *web* por meio de requisições HTTP (Hyper Text Transfer Protocol). No caso específico do comércio eletrônico utilizado, as requisições retornaram JSON (JavaScript Object Notation), mas é possível retornar dados do tipo HTML (Hyper Text Markup Language) e XML (eXtensible Markup Language).

Após a coleta dos dados, foi necessário guardá-los fisicamente. Nesta etapa utilizou-se a estrutura de Data Frame do Pandas, pois ele permite criar uma estrutura tabular. Na Figura 1 é apresentado parte do Data Frame criado, onde são exibidos os comentários originais obtidos e a quantidade de estrelas atribuída pelos usuários.

	Unnamed: 0	comentarios	estrelas
0	0	Comprei o produto, que não veio sequer embalad...	1
1	1	Amei, minha multifuncional, coube direitinho n...	5
2	2	Boa qualidade de impressão e de digitalização ...	5
3	3	Comprei a multifuncional por um valor acessíve...	5
4	4	Otima impressora, recomendado exelente custo b...	5

Figura 1. Data Frame contendo os comentários e a quantidade de estrelas do produto.
Fonte: Elaborado pelos autores

2.2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma subárea da Ciência da Computação, Inteligência Artificial e da Linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais, tendo sido iniciada na década de 1950, quando Alan Turing publicou o artigo "Computing Machinery and Intelligence", que propunha o que agora é chamado de Teste de Turing como critério de inteligência. O Teste de Turing refere-se à habilidade de criar sistemas inteligentes capazes de se relacionarem com pessoas sem que estas saibam que estão se comunicando com máquinas. Nos dias atuais esta área está em destaque principalmente pelo uso de Chatbots no atendimento a clientes.

No pré-processamento realizou-se a normalização semântica das sentenças, ou seja, foram retiradas todas as pontuações e *stopwords*, pois isso pode causar alterações no resultado. Inicialmente, tem-se a sentença original: “O produto é ótimo, compacta, bonita, estou satisfeito”.

Durante o pré-processamento retiraram-se todos os acentos, em seguida foram removidas todas as pontuações e, por fim, removeu-se as *stopwords*, ou seja, palavras que não possuem um significado semântico relevante para a análise, tendo como resultado: “produto otimo compacta bonita satisfeito”.

Com o pré-processamento efetuado, o próximo passo foi converter as sentenças em um vetor de palavras utilizando uma representação numérica conhecida por Bag of Words (BOW), o qual enumera a frequência de cada palavra que apareceu no texto e coloca esses resultados em um vetor.

2.3. Aprendizado de Máquina

O Aprendizado de Máquina é utilizado para extrair padrões complexos de grandes conjuntos de dados. Frequentemente, o objetivo do Aprendizado de Máquina é prever uma determinada resposta com base em uma ou mais variáveis preditoras. Existem muitos modelos de Aprendizado de Máquina, abrangendo métodos básicos, como regressão linear, regressão logística e modelos de árvore, bem como métodos mais sofisticados, como redes neurais artificiais ou máquinas de vetores de suporte (HASTIE *et al*., 2009).

Normalmente, o Aprendizado de Máquina não restringe a análise de dados a apenas um modelo, mas compara muitos modelos e escolhe aquele que atinge a melhor predição. Um desafio comum na análise de dados é a presença de padrões não lineares nos dados, como curvas e interações entre variáveis preditoras. Neste trabalho foi utilizado o método de Regressão Logística.

Entende-se por Regressão Logística uma instância da técnica de classificação que pode ser usado para prever uma resposta qualitativa de acordo com uma variável binária. Um dos benefícios desse método é que os resultados das análises ficam contidos entre um intervalo de 0 e 1 (CABRAL, 2013), que no caso de variáveis categóricas seria o intervalo ideal.

Para rotular o comentário do produto como sendo negativo ou positivo foi necessário utilizar a quantidade de estrelas atribuídas pelo usuário (1 até 5). Neste projeto, considerou-se que um comentário é dito como positivo se a quantidade de estrelas for maior ou igual a 3, e negativo quando possuir uma quantidade de estrelas menor que 3. Estes novos valores foram atribuídos ao Data Frame como uma nova coluna “rótulo”. Por fim, concentrando em uma configuração equilibrada para os experimentos, foi selecionado aleatoriamente 51 comentários positivos e 51 comentários negativos para formar o *corpus* de treinamento e teste.

O classificador binário utilizado foi o de Regressão Logística, escolhido por apresentar melhores resultados quando comparado aos algoritmos de Aprendizado de Máquina supervisionado Árvore de Decisão e Naive Bayes. O classificador tem com entrada um conjunto de dados para treino e um conjunto de dados para teste, pelos quais são possíveis calcular as métricas para avaliação, quais sejam: Medida F1, Acurácia, Precisão e Revogação.

3. Resultados e Discussões

Para esta etapa foram definidos dois experimentos: No primeiro, do total de dados disponíveis, foram utilizados 90% para o treinamento e 10% para testes. No segundo experimento utilizou-se o método de validação cruzada.

3.1 Experimento 1 (90/10)

Na Tabela 1 são apresentados os resultados obtidos no primeiro experimento. Os resultados para todas as métricas de análise utilizadas foram acima de 90%, ajustando-se muito bem ao conjunto de dados anteriormente observado, mas verificou-se que isso aconteceu pois o algoritmo de Aprendizado de Máquina memorizou os dados ao invés de aprender com eles, o que é conhecido por sobreajuste (*overfitting*), e que mostra a ineficácia para prever novos resultados.

Tabela 1- Resultados do Experimento 1

Medida F1	**Acurácia**	**Precisão**	**Revogação**
94,7%	90%	90%	100%

Fonte: Elaborada pelos autores

3.2 Experimento 2 (10-*fold cross validation*)

Na Tabela 2 são apresentados as médias e o desvio padrão (entre parênteses) de cada métrica a partir da validação cruzada utilizando 10 separações, ou seja, o *dataset* foi dividido em 10 partes e estas foram, de maneira aleatória, utilizadas para treinamento e teste. Com isso foi evitado que o algoritmo de Aprendizado de Máquina memorize a base de dados ao invés de aprender com ela. É possível verificar que, ainda assim, foram obtidos resultados satisfatórios, com 89% de precisão.

Tabela 2- Resultados do Experimento 2

Medida F	**Acurácia**	**Precisão**	**Revogação**
89% (+/- 17%)	85% (+/- 26%)	89% (+/- 17%)	98% (+/- 12%)

Fonte: Elaborada pelos autores

3.3 Matriz de Confusão

As métricas apresentadas anteriormente (Medida F1, Acurácia, Precisão e Revogação) foram calculadas a partir da matriz de confusão, na qual são dadas as frequências de classificação para cada classe do modelo. Com isso é possível ver claramente quantos verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos o modelo conseguiu prever. Entende-se por: verdadeiros positivos todos que estamos buscando e foram classificados corretamente; falsos positivos todos que estamos buscando porém foram classificados incorretamente, verdadeiros negativos todos aqueles que não estamos buscando e foram classificados corretamente e, por fim, o falso negativo sendo todos aqueles que não estamos buscando e foram classificados incorretamente.

Analisando a matriz de confusão na Figura 2, tem-se que dos 51 comentários positivos, o sistema reconheceu 41 corretamente e 10 incorretamente. Por outro lado, considerando os 51 comentários negativos, o sistema reconheceu 50 corretamente e 1 incorretamente.

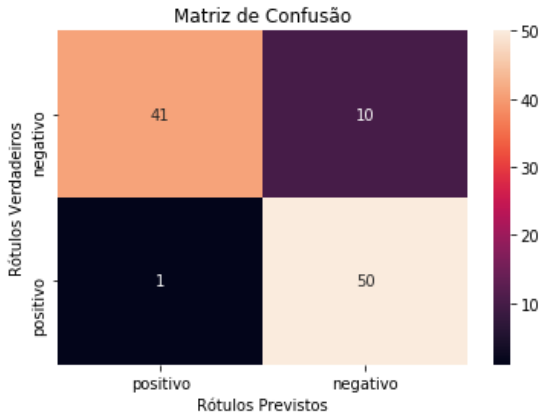

Figura 2. Matriz de Confusão
Fonte: Elaborado pelos autores

4. Conclusão e Trabalhos Futuros

A partir desta pesquisa foi possível verificar como diferentes abordagens de treinamento e teste de Aprendizado de Máquina supervisionado tradicional se comparam, em termos de desempenho, na mineração de opiniões, ou seja, na classificação binária dos comentários como sendo positivos ou negativos.

Conclui-se que o sistema, mesmo com um *dataset* relativamente pequeno, pode gerar resultados satisfatórios em ambos os casos, porém a técnica de validação cruzada apresentou melhor resultado para o problema proposto dada a eliminação de sobreajuste.

Como trabalhos futuros pretende-se realizar a captura de novas avaliações para o *corpus*, assim como realizar o estudo e aplicações de outras técnicas de Processamento de Linguagem Natural (*Word Embeddings*) e Aprendizado de Máquina para melhorar os resultados na mineração de opiniões.

Referência Bibliográfica

ANDRADE, Maria Margarida. Como preparar trabalhos para cursos de pós-graduação: noções práticas. 5. ed. São Paulo: Atlas, 2002.

CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. 2013. Dissertação de Mestrado em Matemática Aplicada à Economia e à Gestão apresentado ao Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa, Lisboa, 2013. Disponível em: http://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 29 maio 2020.

GUILHARDI, H. J. Análise comportamental do sentimento de culpa. In: Ciência do comportamento: conhecer e avançar, v.1, p. 173–200, 2002.

HASTIE, T.; TIBSHIRANI R.; FIREDMAN, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

HORRIGAN, J. *Online Shopping*. Pew Internet and American Life Project Report Google Scholar, 2008.

KOZINETS, Robert. *The field behind the screen: Using the method of netnography to research market-oriented virtual communities*, 2001.

LIU, B. *Sentiment Analysis And Subjectivity*. Handbook of Natural Language Processing, 2010.

MEDHAT, W.; HASSAN, A; KORASHY, H. *Sentiment analysis algorithms and applications: a survey*. In: Shams Engineering Journal. v.5 (4), p.1093-1113, 2014.

MEULEMAN, B; SCHERER, K. R. *Nonlinear appraisal modeling: An application of machine learning to the study of emotion production*. In: IEEE Transactions on Affective Computing, v. 4 (4), p. 398-411, 2013.

PANG, B.; LEE, L. *Opinion mining and sentiment analysis*. In: Foundations and Trends in Information Retrieval, v.2 (1), p.1-135, 2008.

SKINNER, B. F. Questões recentes na análise comportamental. [S.l.]: Papirus, 1991.

SOUZA, M. V. S. Mineração de opiniões aplicada a mídias sociais. Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, 2012.