Universidade Federal do Tocantins - Campus Palmas
Ciência da Computação/9º Período
Aprendizado de Máquinas
Prof. Carlos Rodrigues
Ítalo Machado Vilarino

Relatório: Classificação de Sentimentos em textos utilizando Aprendizado de Máquina

1. INTRODUÇÃO

O campo do aprendizado de máquina tem ganhado cada vez mais destaque nas últimas décadas, impulsionado pelo crescente volume de dados e avanços na capacidade computacional. Essa área da inteligência artificial visa desenvolver algoritmos capazes de aprender padrões e tomar decisões a partir dos dados, sem a necessidade de uma programação explícita.

Uma das tarefas mais comuns no aprendizado de máquina é a classificação, que determina a atribuição de rótulos ou categorias a dados com base em suas características. A classificação é aplicada em uma ampla gama de domínios, desde diagnósticos médicos até análise de sentimento em mídias sociais.

A análise de sentimento é uma das aplicações mais relevantes e desafiadoras da classificação. Ela envolve a identificação e categorização do sentimento expresso em um texto, seja positivo, negativo ou neutro. A capacidade de compreender o sentimento dos usuários em relação a um produto, serviço ou evento pode fornecer estratégias valiosas para empresas e pesquisadores.

Este trabalho dedica-se à classificação de sentimentos utilizando técnicas de aprendizado de máquina. A partir de um conjunto de dados que contém textos acompanhados de avaliações, buscou-se desenvolver um modelo capaz de prever automaticamente o sentimento associado a um texto desconhecido. Para isso, foram aplicadas técnicas de pré-processamento, alinhamento de dados e também alguns algoritmos de classificação específicos para análise de sentimentos.

Ao longo deste relatório, serão apresentados os detalhes das etapas do processo, incluindo as técnicas de pré-processamento utilizadas, a seleção e treinamento dos algoritmos de classificação, além da avaliação e interpretação dos resultados obtidos.

O objetivo principal deste trabalho será explorar a capacidade do aprendizado de máquina em classificar sentimentos em textos, fornecendo uma visão abrangente das técnicas e abordagens utilizadas.

2. METODOLOGIA

2.1. Definição do Problema

Para iniciar o projeto foi definido o problema que seria abordado no presente trabalho, que consiste na classificação de sentimentos em textos. Identificou-se a importância dessa tarefa em diferentes domínios e o escopo do problema foi delimitado dentro do dataset "sentiment-emotion-labelled_Dell_tweets.csv" para garantir um foco adequado.

2.2. Pré-processamento dos dados

Definido o escopo de trabalho, foi realizado o pré-processamento destes dados, que incluiu as seguintes etapas:

- Limpeza dos textos: remoção de caracteres especiais, pontuações e números indesejados.
- Conversão para letras minúsculas: padronização de todas as palavras em minúsculas para evitar diferenciação entre maiúsculas e minúsculas.
- Remoção de stopwords: eliminação de palavras comuns (na língua inglesa) que não contribuem para a classificação de sentimentos, como artigos, preposições e pronomes.
- Lematização: redução das palavras à sua forma básica, para evitar a variação morfológica e reduzir o nosso escopo.
- Stemming: redução das palavras à sua raiz, para eliminar sufixos e prefixos e simplificar a representação textual.

2.3. Representação dos Dados

Nesta etapa, foi realizada a representação dos dados pré-processados em uma forma numérica que pudesse ser utilizada pelos algoritmos de aprendizado de máquina. O

método de Term Frequency-Inverse Document Frequency (TF-IDF) foi utilizado para atribuir importância às palavras com base em sua frequência nos textos e no conjunto de dados como um todo.

2.4. Seleção e Treinamento dos Algoritmos de Classificação

Neste módulo, foram selecionados alguns algoritmos de classificação, de forma a abranger um pouco de tudo que foi visto no escopo da disciplina. Para alcançar um resultado satisfatório para o problema de classificação de sentimentos em textos, utilizando a metodologia de pré-processamento que foi abordada acima e mantendo o TF-IDF para representação dos dados, os testes foram realizados em quatro abordagens tradicionais: Logistic Regression, Naive Bayes, Support Vector Machines (SVM) e Árvores de Decisão. Essa seleção foi realizada sem um estudo profundo dos dados selecionados ou considerações sobre a compatibilidade com os métodos de pré-processamento utilizados. Os testes foram realizados considerando única e exclusivamente a capacidade que estes algoritmos possuem de lidar com dados textuais e em sua eficácia em problemas de classificação de sentimentos em geral.

Os conjuntos de dados foram divididos em conjuntos de treinamento e teste e foram treinados os algoritmos selecionados com os dados de treinamento.

Linguagem e Tecnologias Utilizadas:

- Linguagem de programação: Python
- Bibliotecas: scikit-learn, NLTK (Natural Language Toolkit), pandas, numpy
- Ferramentas de pré-processamento de texto: stopwords, WordNetLemmatizer, PorterStemmer
- Algoritmos de classificação: Naive Bayes, Support Vector Machines (SVM), Árvores de Decisão, Redes Neurais Artificiais
- Técnicas de representação de dados: TF-IDF
- Dataset: kaggle.com/datasets/ankitkumar2635/sentiment-and-emotions-of-tweets

3. RESULTADOS

Neste trabalho, foram aplicados diferentes algoritmos de classificação para a tarefa de análise de sentimentos em textos. O objetivo deste trabalho era determinar qual algoritmo obteve o melhor desempenho na classificação de sentimentos com base em uma predisposição de dados comum.

Após realizar a avaliação dos algoritmos em um conjunto de dados específico, foram obtidos os seguintes resultados de precisão:

- Decision Tree Classifier: Precisão de 0.644973968762515
- SVM: Precisão de 0.7615138165798959
- Naive Bayes: Precisão de 0.6774128954745695
- Logistic Regression: Precisão de 0.76431718061674

Com base nos resultados, é possível observar que o algoritmo com a maior precisão foi o Logistic Regression, alcançando um valor de 0.76431718061674. Em seguida, o SVM obteve uma precisão de 0.7615138165798959. O Naive Bayes apresentou uma precisão de 0.6774128954745695, enquanto o Decision Tree Classifier teve o menor desempenho, com uma precisão de 0.644973968762515.

Tais resultados podem ser úteis para avaliações e métricas que sejam similares ao problema apresentado. No entanto, é importante ressaltar que não existe um algoritmo ideal e sua escolha sempre vai depender do conjunto de dados e das características do problema. Portanto, é recomendado realizar testes adicionais e considerar outras métricas de avaliação para uma análise completa do problema.