Projeto de Inteligência Artificial

Kevin Mayan, Marcelo Freitas, Samael Aun

3 de julho de 2024

Resumo

Este artigo descreve a construção e avaliação de pipelines para a análise de sentimentos e detecção de toxicidade em tweets. Usando técnicas de processamento de linguagem natural (PLN) e algoritmos de aprendizado de máquina, alcançamos métricas de precisão, recall e f1-score satisfatórias. Os resultados mostram a eficácia das abordagens propostas em identificar sentimentos e conteúdos tóxicos em textos curtos.

1 Introdução

A análise de sentimentos e a detecção de toxicidade em textos online são importantes para diversas aplicações, como monitoramento de redes sociais, moderação de conteúdo e melhoria da experiência do usuário. Este trabalho apresenta duas abordagens distintas para esses problemas, utilizando técnicas de PLN e algoritmos de aprendizado de máquina.

2 Metodologia

Realizamos o pré-processamento, limpeza de texto, remoção de stopwords, lematização dos dados e representamos as palavras através da vetorização TF-IDF. Para a classificação, foi utilizado o algoritmo Naive Bayes. Métricas de avaliação utilizadas: acurácia, precisão, recall e f1-score.

3 Trabalhos Relacionados

A análise de sentimentos e a detecção de toxicidade têm sido amplamente estudadas na literatura. Técnicas comuns incluem o uso de vetorização TF-IDF, embeddings de palavras como Word2Vec e algoritmos de classificação como Naive Bayes, SVM e redes neurais. Trabalhos recentes também exploram o uso de transformers, como BERT, para melhorar o desempenho dessas tarefas.

3.1 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF (Term Frequency-Inverse Document Frequency) é uma técnica amplamente utilizada em processamento de linguagem natural (PLN) e recuperação de informações para avaliar a importância de uma palavra em um documento em relação a um corpus. É uma combinação de duas métricas: 1. **Term Frequency (TF):** Mede a frequência de uma palavra em um documento específico. 2. **Inverse Document Frequency (IDF):** Mede a raridade da palavra em todo o corpus de documentos.

A fórmula do TF-IDF é:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

onde: - t é o termo (palavra). - d é um documento específico. - D é o corpus de documentos.

A importância do TF-IDF reside na sua capacidade de reduzir a importância de palavras comuns (como artigos e preposições) e aumentar a importância de palavras raras que podem ser mais indicativas do conteúdo de um documento. Isso é particularmente útil para tarefas de análise de sentimentos e detecção de toxicidade, onde palavras específicas podem ser fortes indicadores de sentimentos ou de toxicidade.

3.2 Naive Bayes

O algoritmo Naive Bayes é um classificador probabilístico baseado no teorema de Bayes com a suposição de independência entre os atributos. Ele é chamado de "ingênuo"porque assume que todas as características são independentes umas das outras, o que raramente é verdade na prática, mas ainda assim funciona bem em muitos cenários.

O teorema de Bayes é definido como:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

onde: - $P(C|X)$ é a probabilidade posterior da classe C dado o atributo X. - $P(X|C)$ é a probabilidade de X dado que C é verdadeiro. - $P(C)$ é a probabilidade anterior da classe C. - $P(X)$ é a probabilidade anterior do atributo X.

Em PLN, o Naive Bayes é frequentemente usado para classificação de texto, como na análise de sentimentos e na detecção de spam, devido à sua simplicidade e eficiência. Ele é particularmente eficaz quando aplicado a conjuntos de dados grandes e é capaz de lidar com alta dimensionalidade, uma característica comum em dados textuais.

Estas referências fornecem uma base sólida para entender as técnicas de TF-IDF e Naive Bayes, que são fundamentais para a análise de sentimentos e a detecção de toxicidade em textos. Além disso, elas destacam a importância e a aplicação dessas técnicas em cenários de PLN.

4 Resultados

Métrica	**Sentimento**	**Toxicidade**
Acurácia	0.572	0.918
Precisão	0.717	0.908
Recall	0.572	0.902
F1-score	0.512	0.905

Tabela 1: Métricas de desempenho para análise de sentimentos e detecção de toxicidade

5 Discussão

Os resultados obtidos demonstram a eficácia do algoritmo Naive Bayes combinado com a vetorização TF-IDF para tarefas de análise de sentimentos e detecção de toxicidade. Embora a acurácia na análise de sentimentos (0.572) seja inferior à obtida na detecção de toxicidade (0.918), isso pode ser explicado pela natureza mais subjetiva e variada dos sentimentos expressos em textos curtos, como tweets. A alta acurácia na detecção de toxicidade indica que o modelo conseguiu identificar padrões claros de linguagem associada a conteúdo tóxico.

Além disso, a precisão (0.717) e o recall (0.572) na análise de sentimentos sugerem que o modelo é mais eficaz em identificar verdadeiros positivos do que em evitar falsos negativos. Isso implica que, embora o modelo consiga detectar muitos exemplos de sentimentos corretos, ele também perde uma quantidade significativa deles. Em contraste, os resultados de precisão (0.908) e recall (0.902) na detecção de toxicidade mostram um desempenho equilibrado, o que é crucial para aplicações práticas onde a identificação correta de conteúdos tóxicos é essencial para a moderação de conteúdo.

6 Conclusão

Neste trabalho, apresentamos a construção e avaliação de pipelines para a análise de sentimentos e detecção de toxicidade em tweets, utilizando técnicas de processamento de linguagem natural e algoritmos de aprendizado de máquina. Nossos resultados indicam que o Naive Bayes, quando combinado com a vetorização TF-IDF, é uma abordagem viável para essas tarefas, alcançando métricas de desempenho satisfatórias.

Para trabalhos futuros, sugerimos explorar técnicas de deep learning, como transformers e BERT, que têm demonstrado melhorar significativamente o desempenho em tarefas de PLN. Além disso, investigar a combinação de múltiplos algoritmos de classificação e técnicas de ensemble pode levar a melhorias adicionais na precisão e robustez dos modelos.

Referências