Análise de Sentimentos em Textos

Jean Luckei Tolotti

Universidade do Vale do Rio dos Sinos (Unisinos)

Resumo. *Este artigo descreve o processo de escolha de métodos de Machine Learning a partir de um conjunto de dados coletados através da plataforma Twitter. Contempla a metodologia, o processamento dos dados, as análises executadas e os modelos obtidos, para ao fim gerar insumos para um algoritmo de predição de sentimentos baseado nas palavras obtidas da plataforma em questão, sendo possível identificar sentimentos positivos ou negativos de acordo com um conjunto de palavras aleatórias.*

1. Informações Gerais

Este relatório tem por objetivo explanar o problema identificado, utilizar algum algoritmo de Machine Learning para resolvê-lo e então demonstrar suas métricas e aceitação.

O problema encontrado é a incapacidade da máquina de interpretar emoções e então definir o que o escritor está expressando. Se um texto informado pode ser considerado sentimentalmente negativo ou positivo, cabe ao leitor entender, coisa que a máquina transmissora desta mensagem é incapaz de fazer.

Com isso, surge a necessidade de ensinar a máquina, utilizando de modelos de Machine Learning, como interpretar sentimentos positivos ou negativos, a fim de entender o texto publicado em comentários de fotos, produtos, páginas da internet, fóruns etc. Os fins dessa interpretação podem ser variados, desde sentimento quanto à marca que faz algum anúncio até análises de tendências em certos grupos.

2. O Conjunto de Dados

Para treinar uma máquina, é necessário um dicionário de dados muito extenso e válido, pois a partir dele serão gerados os vetores capazes de definir pesos para as palavras contidas em uma frase, o que acaba por definir se uma frase é negativa ou positiva.

O Dataset encontrado foi o “sentiment140” que está disponível no seguinte endereço web https://www.kaggle.com/kazanova/sentiment140 e possui 1.6Mi de registros de postagens extraídas da plataforma Twitter. Com este número de mensagens de texto é possível gerar muitas combinações de palavras, e como este dataset já possui valores de sentimento previamente definidos por humanos, ele é de grande confiabilidade.

Este conjunto de dados contém duas colunas que serão objeto de estudo deste relatório, sendo uma delas a coluna “target” (chamada de sentiment neste trabalho), que descreve o sentimento em positivo (4) ou negativo (0), e a coluna “text” que possui o texto do tweet.

3. Pré-Processamento

Os dados disponibilizados no dataset precisaram sofrer alguns ajustes. Como podemos notar o atributo “sentiment” pode ser definido como um binário (aumentando as chances de utilizarmos uma técnica de classificação, pois o alvo é sentimento positivo ou negativo, portanto 0 ou 1) e por isso podemos alterá-lo de 0 e 4 para 0 e 1, apenas realizando o replace de todos os valores 4 por 1.

Além deste processamento, foi necessário um mais complexo no atributo “text”, visto que o fator humano está envolvido, é possível encontrar erros gramaticais, vícios de linguagem, uso excessivo de emojis e palavras desnecessárias para os algoritmos de Natural Language Processing (NLP).

Para isso, foi realizada a transformação de todos os textos em lowercase (minúsculo), foi feita a substituição de URLs iniciadas por "http", "https" e "www" para o coringa “URL”, foi feita a substituição dos @nomeusuario por "USER", a remoção de caracteres que não são alfabéticos, além da remoção de palavras com menos de 2 letras e palavras com a mesma letra repetida muitas vezes.

Foram removidas as palavras irrelevantes, chamadas de “stopwords”, utilizando-se da biblioteca nltk. Com esta mesma biblioteca foi possível aplicar a lematização, que tem como objetivo reduzir uma palavra à sua forma base e agrupar diferentes formas da mesma palavra.

4. Análises e Modelos

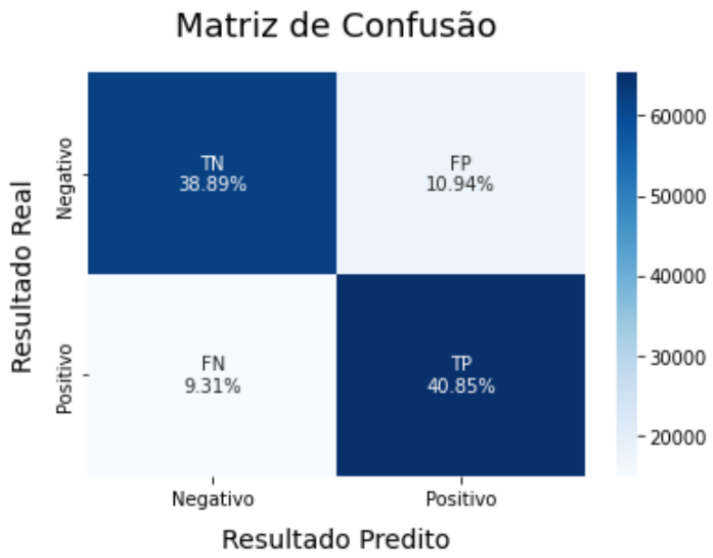
Como foi identificado que os dados possuem uma variável alvo binária e que outras variáveis dentro do texto podem determinar o seu valor, optou-se pelos modelos de classificação, neste caso a regressão logística, tendo como variável dependente Y o conjunto de dados da coluna “sentiment”, e o conjunto de dados gerado pelo processamento dos textos sendo o nosso f(X) que deverá treinar o algoritmo para ter a capacidade de inferir o Y.

Tendo isso definido, precisamos criar o nosso X de uma forma mais inteligente, aplicando vetorização a partir da matriz de TF-IDF, a qual é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos, gerando assim pesos diferentes para cada palavra em uma frase, o que será fator determinante no valor do Y.

Utilizando as bibliotecas disponíveis no python, mais especificamente a função vectoriser.transform(x_train), foi possível chegar a uma matriz de treino e outra de teste para o X, as quais serão utilizadas no modelo de regressão logística.

O resultado da acurácia ficou em 80%, o que consideramos ser o suficiente para identificação de textos num contexto não crítico. Com um dicionário maior seria possível obter melhores resultados.

Abaixo é demonstrada a matriz de confusão:



Pode-se calcular também a precisão do modelo com a fórmula $TP/(TP+FP)$, chegando ao resultado de ~0.79.

4. Conclusão

Em suma, o modelo respondeu muito bem aos testes, apesar de apenas 80% de acurácia no teste vs treino que estava definido em 10% para 90%, concluímos que o algoritmo é bem aceito para predições de sentimentos em texto. Exemplo de testes executados:

```
# Texto de teste
text = ["When are you inviting us to your place? It's been so long.",
        "Today is not going to be a good day, sorry",
        "@testuser, I'd like to send you a happy new year message.",
        "The president is dead, my condolences to his family and friends.",
        "Tomorrow is going to be the best day of your life, as always! :)",
        "Please, don't talk to me anymore..."]

# Gera um dataframe com os resultados
df = predict(vectoriser, model, text)
print(df)
```

```
                                               Texto Sentimento
0  When are you inviting us to your place? It's b...   Negativo
1         Today is not going to be a good day, sorry   Negativo
2  @testuser, I'd like to send you a happy new ye...   Positivo
3  The president is dead, my condolences to his f...   Negativo
4  Tomorrow is going to be the best day of your l...   Positivo
5                Please, don't talk to me anymore...   Negativo
```

O próximo objetivo deste estudo será adquirir um dataset com pontuações para textos na língua portuguesa, para então reproduzir os mesmos treinamentos e identificar se os métodos de tratamento também serão aplicáveis a outra língua.