Relatório TP4-ORI

Laura Rosado Rodrigues Muniz 12021BSI216

Sara Rosado Rodrigues Muniz 12021BSI215

1. Ao utilizar TF-IDF e 1-gram em vez do CountVectorizer, passamos a representar os dados de texto com base na frequência do termo (TF) e na frequência inversa (IDF) do documento, o que pode proporcionar melhores resultados em comparação com a contagem simples da frequência do termo. Por outro lado, quando comparamos o primeiro experimento com o segundo, é possível perceber que os valores de precision, recall e f1-score diminuíram e a quantidade de acertos por classe também diminuiu.

ANTES

Calendário

Descrição gerada automaticamente

DEPOIS

Calendário

Descrição gerada automaticamente

1. Ao comparar o 1-gram e o 2-gram é possível perceber que houve uma piora quanto a classificação. Um exemplo disso seria que ao analisarmos o número de acertos por classe, por meio da matriz de confusão, é possível visualizar, por exemplo, que o número de classificações negativas que foram classificadas como positivas de forma erronia aumentou de 22 para 26. Por outro lado, a acurácia manteve em 0.88.

1-Gram:

Calendário

Descrição gerada automaticamente

2-Gram:

Calendário

Descrição gerada automaticamente

1. Ao compararmos todas as quatro possibilidades é possível perceber que todas as acurácias são de 0.88. Por outro lado, os valores de precision, recall, f1-score e o número de acertos por classe são diferentes para cada possibilidade. Assim, é possível concluir que o modelo TF foi melhor que os outros, pois seus valores de precision, recall e f1-score foram razoavelmente maiores, o que significa que a capacidade do modelo de identificar corretamente os exemplos é melhor e o equilíbrio entre precision e recall também. Além disso, o número de erros ao identificar as classificações foi menor quando comparado com os demais modelos.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamenteTabela

Descrição gerada automaticamente com confiança média

Comparando com os 3 algoritimos adicionados (Árvore de Decisão, Random Florest e SVM) o melhor resultado foi o do algorítimo SVM, pois o valor de sua acurácia foi o maior totalizando 0.90. Além disso, seu número de acertos por classe foi maior e seus valores de precision, recall e f1-score também foram razoavelmente maiores quando comparado com os demais.

Calendário

Descrição gerada automaticamente

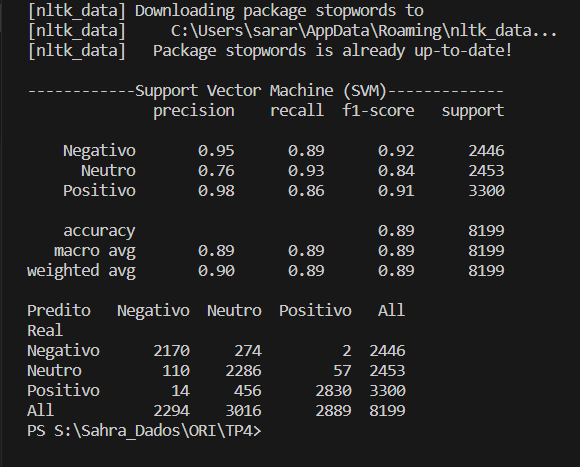
Calendário

Descrição gerada automaticamente

Calendário

Descrição gerada automaticamente

1. Ao utilizar o algoritmo SVM com o CountVectorazer (configuração que apresentou melhor resultado) e utilizar a biblioteca nltk para remover as stopwords, a acurácia do modelo diminuiu para 0.89, juntamente com o número de acertos por classe.



1. Quando comparamos os resultados da avaliação do arquivo “reforma\_previdencia\_rotulado.csv” com os resultados retornados pelo arquivo “Tweets\_Mg.csv”, percebe-se que a acurácia do modelo caiu drasticamente de 0.89 para 0.60, de modo a gerar, consequentemente, mais erros na classificação entre “positivo”, “negativo” e “neutro”. Logo, os resultados entre os arquivos não são semelhantes. Tal fator pode ocorrer por vários motivos, entre eles estão a variabilidade, diferenças na composição, tamanho e qualidade dos conjuntos de dados.

Calendário

Descrição gerada automaticamente