**Relatório Trabalho Final:**

**Análise de Sentimento de “*Tweets*” Referentes a Cantora Anita**

Aluno: Claudionor Ferreira da Silva Junior

Disciplina: Aprendizado de Máquinas I

Professor: Douglas Rodrigues

**Contexto Geral:**

Anitta é uma cantora, compositora e empresária brasileira conhecida por misturar pop, funk carioca e reggaeton. Além de sua música, ela é uma influente figura nas redes sociais, reconhecida por sua versatilidade e impacto cultural. Este estudo analisou publicações de usuários (ou “*tweets*”) referentes à cantora na plataforma Twitter (atual “X”) para construir uma análise de sentimento e um modelo que classifica esses tweets como positivos ou negativos em relação à celebridade.

**Constituição da Base de Dados Original:**

A base de dados é constituída pelas colunas:

* **user\_id –** Identificação do autor da publicação
* **conversation\_id –** Identificação da conversa da publicação
* **user\_username –** Nome de usuário do autor da publicação
* **text –** Publicação referente a cantora
* **Polaridade –** Identificação se o “*tweet*” é bom ou ruim (ruim = -1 e bom = 1)

**Etapas do Pré-Processamento:**

1. Primeiramente, foi realizada uma divisão da base de dados em treino e teste utilizando a biblioteca **“*caret*”**, na qual todos os pré-processamentos abaixo foram aplicados primeiramente na base treino para fazer o treinamento do modelo, e posteriormente, será aplicado na base teste para verificação das métricas de Acurácia, Sensibilidade e Especificidade.
2. Após essa separação, foram retiradas as colunas “user\_id”, “conversation\_id” e “user\_username”, utilizando a biblioteca **“*dplyr*”,** para só trabalharmos com as colunas das publicações e identificação.
3. Nas publicações, foram feitas manipulações para adequação e melhora de desempenho na análise de sentimento e classificação, como a remoção de acentos das palavras, remoção de números e substituição de emojis por palavras que realmente representam o sentimento. Além disso, foi criada uma coluna “RecordID” para identificação de cada publicação. Tudo isso utilizando a biblioteca **“*dplyr*”**
4. Foi realizado a **tokenização** da coluna de “*tweets*”, ou seja, as frases foram separadas em palavras. Além disso, para retirar as chamadas **“*stopwords*”,** ou palavras que não agregam na análise de sentimento, como conjunções e conectivos, foi usado um dicionário que contém palavras acrescentadas por mim: <https://raw.githubusercontent.com/Claudionor20/XSentiment/main/stopwords_.txt>
5. Nas publicações há uma grande variabilidade de “gírias” ou formas diferentes de falar uma palavra, como por exemplo, a palavra “arrasar” foi escrita como “arrasa” ou “arrasouuu”. Com isso, foi feito uma manipulação para converter esse tipo de variações a uma só palavra
6. Nessa etapa foi feita a chamada “lematização”, ou seja, a transformação de palavras similares por uma só, na maioria por seu infinitivo ou sinônimo. Para isso, foi usado um dicionário com as palavras e seus respectivos sinônimos, também acrescentado por mim: <https://raw.githubusercontent.com/Claudionor20/XSentiment/main/lematizacao_claudio_v2.txt>
7. Nesse momento foi feita a manipulação com a frequência das palavras, na qual foi retirada as palavras que possuem uma frequência menor que 15 na base de referência. Além disso, para a formação da matriz de termo documento, ou seja, a matriz com a frequência das palavras em cada publicação, foi calculado a frequência percentual de cada palavra em cada publicação. Tudo isso utilizando o pacote **“*dplyr***” e para a criação da matriz termo-documento, foi utilizado a biblioteca **“maditr”**
8. Foram retiradas algumas palavras desnecessárias, como verbos genéricos e palavras que possuem alta correlação entre si. Para detecção da correlação das palavras, foi usado a função a “*findCorrelation”* do pacote**“*caret*”** e estabelecido um parâmetro de exclusão de palavras que possuem uma correlação acima 0.75 entre alguma variável explicativa. Dessa forma, foram excluídas as colunas:

**"V1" , "l", "achar" , "acreditar", "acontecer", "acompanhar", "antiga", "bbma", "chegar", "caro", "causar", "esquerdo"** **, "lindo", "melhorar", "menino", "rede", "brunamarquezine", "esquerda" e "redar"**

1. **ESCREVER SOBRE BALANCEAMENTO**

**Treinamento do Modelo:**

Após concluir o pré-processamento da base de dados de treino, deu-se início ao treinamento do modelo de classificação para identificar pessoas com possibilidade de AVC daqui a 10 anos. Para isso, foi feito validação cruzada a fim de testar os melhores parâmetros para o modelo de “gradient boosting”, usando k = 10. O modelo que melhor desempenhou tinha como parâmetros:

n.tree = 500 ; interaction.depth = 3 ; shrinkage = 0.01 ; n.minobsinnode = 5

Além disso, foram atribuídos pesos as classificações da variável resposta no treinamento do modelo, a fim de corrigir o conjunto de dados desbalanceado, já que, após uma análise descritiva foi visto um número muito grande de pessoas sem tendência a ter AVC em relação as pessoas com tendência. O peso atribuído para cada classe foi a razão do número de amostras totais e o número de amostras de cada classificação.

Por fim, o ponto de corte da classificação do risco de AVC foi feito determinado utilizando-se curva ROC.

**Resultados do Modelo:**

Para avaliar a qualidade do modelo, empregou-se a base de dados de teste, que havia sido separada no início do estudo, aplicou-se o mesmo pré-processamento utilizado na base de treino e foi feito uma matriz de confusão a partir do modelo treinado. Os resultados obtidos foram os seguintes:

|  |  |  |  |
| --- | --- | --- | --- |
| Medida | Acurácia | Sensibilidade | Especificidade |
| Valor (%) | 73.94% | 75.91% | 63.58% |