PUCRS – Escola Politécnica
Ciência de Dados e Inteligência Artificial – 2022/2

Trabalho 2: Inteligência Artificial

Análise de Sentimentos – Machine Learning

Lorenzo Lazzarotto - 21101016
Thiago Macedo - 21104690
Pedro Nunes - 20110026
Vitor Pires - 21101680

Porto Alegre – RS

Introdução

Este relatório tem como objetivo descrever o trabalho 2 da disciplina de Inteligência Artificial, explicando o problema que foi abordado e a solução elaborada, mostrando detalhes de sua implementação em Python e os resultados obtidos.

Enunciado do Trabalho

O trabalho consiste em utilizar machine learning para realizar análise de sentimentos em texto de tweets gerados durante as eleições de 2022 no estado do Rio Grande do Sul. O conjunto de dados continha 3000 tweets não anotados, especificando o id do usuário, o número de RTs e o próprio texto do tweet.

O grupo deveria anotar os dados (0 para um tweet negativo e 1 para positivo), fazer o pré-processamento e criar pelo menos quatro modelos de machine learning para classificar corretamente cada tweet.

Pré-Processamento dos Dados

A etapa de pré-processamento é essencial para a geração de um bom modelo. Portanto, tentamos limpar e preparar ao máximo o conjunto de dados. Ao final do processo, os dados estavam prontos para serem alimentados para os algoritmos. O pré-processamento foi implementado com as bibliotecas Pandas, NLTK e Scikit-Learn, através das seguintes etapas:

1. Remoção das colunas desnecessárias (mantemos somente o tweet e a classificação)
2. Remoção dos registros sem classificação (nulos)
3. Remoção dos valores duplicados (tweets iguais)
4. Transformação do texto dos tweets para minúsculo
5. Remoção dos caracteres especiais de cada tweet
6. Tokenização do texto de cada tweet
7. Lematização do texto de cada tweet
8. Remoção das ‘stop words’ de cada tweet
9. Aplicação do método TF-IDF nos dados

Modelos Utilizados

Após a etapa de pré-processamento, foram utilizados diversos modelos de machine learning, em busca do melhor classificador possível. Todos os modelos foram implementados através da biblioteca Scikit-Learn. Os algoritmos utilizados foram os seguintes:

1. KNN
2. Multilayer Perceptron
3. XGBoost
4. AdaBoost
5. Naive Bayes
6. Support Vector Machine
7. Random Forest
8. Regressão Logística
9. Stacking (MLP + KNN + Regressão Logística)

Resultados

Após a utilização de dos algoritmos listados acima, os seguintes resultados foram obtidos:

Modelo	F1	Acurácia	Precisão	Recall	TP	TN	FP	FN
KNN	0.82	0.82	0.82	0.82	123	60	18	23
MLP	0.85	0.85	0.85	0.85	129	62	12	21
XGB	0.76	0.77	0.77	0.77	125	48	16	35
AB	0.74	0.75	0.75	0.75	122	46	19	37
NB	0.76	0.78	0.81	0.78	138	37	3	46
SVM	0.79	0.81	0.83	0.81	138	43	3	40
RF	0.81	0.82	0.83	0.82	135	49	6	34
RL	0.82	0.83	0.86	0.83	139	47	2	36
Stack	0.84	0.84	0.84	0.84	130	58	11	25

Conclusão

O trabalho 2 da disciplina de Inteligência Artificial buscou explorar todas as etapas de um projeto de ciência de dados, desde a anotação dos dados até o desenvolvimento do modelo. O projeto também está disponível no GitHub.