

X

×

PARA RESPOSTAS NA PLATAFORMA ESIC

Por Felipe M. Castro e Nilton Enta



SUMÁRIO

×

- PRÉ-PROCESSAMENTO
- COMBINAÇÕES
- MODELOS SELECIONADOS
- ENSEMBLE
- RESULTADOS FINAIS

PRÉ-PROCESSAMENTO

CAIXA BAIXA

×

Testo para caixa baixa com str.lower()

LEMMATIZAÇÃO

Usamos 'pt_core_news_sm' da spacy

PONTUAÇÃO

Remoção de pontuação

RÓTULOS

Encoding dos rótulos

×

TOKENIZAÇÃO

Separamos o texto em palavras

VETORIZAÇÃO

Usamos o TFIDF da NLTK

× COMBINAÇÕES

MODELO	VETORIZAÇÃO	HIPERPARÂMETROS	ACURÁCIA
Naive Bayes	TFIDF	Default	45.30%
SVM	TFIDF	Class_weight='balanced', probability=True	48.90%
Random Forest	TFIDF	n_estimators=167, max_depth=176	51.76%
DNN	TFIDF	512 relu, 0.8 dout, 126 relu, 4 stmax	49.02%
LSTM	TFIDF	64 relu, 0.4 dout, 32 relu, 4 stmax	36.59%
LSTM+C	TFIDF	F32 conv1d. 500 mpool, 64 lstm, 0.4 dout, 64 relu, 4 stmax	36.59%
S ∖ M+RF	TFIDF	-	51.28%
Random Forest	Word2Vec	Default	37.86%

MODELOS SELECIONADOS

0.6

0.5

0.4

0.3

0.2

0.1



ENSEMBLE - VOTING CLASSIFIER





Class_weight='balanced', probability=True weights=[3, 2, 2, 1]

n_estimators=167, max_depth=176 weights=[1, 1, 1, 3]



AVALIAÇÃO DO MODELO



Média do cross validation com 10 folds



Acurácia dentro do conjunto de teste (20% dos dados)



ALGUMA DÚVIDA?

×

Repositório: https://github.com/felmateos/age-group-classifier/ Repositório:





slidesgo