

Spam-Classfier

فاطمه خدادادی-97143013

سوال: پیش بینی کنید که آیا SMS, هرزنامه(Spam) است یا خیر؟

استفاده از الگوریتم های مختلفی مثل :

multinomial-naive-bayes,logistic regression,svm,decision trees

برای مقایسه دقت.

و همچنین استفاده از تکنیک های مختلفی برای پاکسازی(data cleaning) و پردازش داده ها مثل PorterStemmer,CountVectorizer,TFIDF Vetorizer,WordnetLemmatizer

از LSTM و Word Embeddings استفاده شده برای رسیدن به دقت 97.84%

Accuracy :

Text Preprocessing Type	Logistic Regression	Multinomial NB	Support Vector Machine	Decision Tree
TFIDF Vectorizer + PorterStemmer	96.68%	97.30%	98.47%	96.68%

Text Preprocessing Type	Logistic Regression	Multinomial NB	Support Vector Machine	Decision Tree
CountVectorizer + PorterStemmer	98.65%	98.56%	98.74%	97.84%
CountVectorizer + WordnetLemmatizer	98.56%	98.29%	98.38%	97.75%
TFIDF Vectorizer + WordnetLemmatizer	96.41%	97.48%	98.47%	96.86%

مراحل کار:

- 1_Dataset(sms spam)
- 2_Pre-processing phase
- 3_Feature selection and extraction
- 4_Classifier training
- 5_Classifier testing
- 6_Classification results
- 7_Performance evaluation