Fake news classification using NLP

Abdullah Al-Huwaishel

Ghanim Al-Ghanim

Abstract

The goal of this project was to use classification to predict if the news is fake or true . We worked with the data from https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset , leveraging feature selection, feature engineering, word cleaning, and tokenize. Then, we built Logistic Regression, KNN, and Random Forest models. We concluded by comparing between the accuracy of each model.

Design

This project originates from kaggle database. Classifying news accurately via machine learning models would classifiy if the input is fake or true. Also, it would enable user to know if the news is accurate or not.

Data

fake news dataset contains 44,000 data points and 4 features for each data point. A few feature highlights include title, text, subject, and date, and our target category.

but we only need title and text and remove the other features since we are gonna do nlp and the other features are not string.

After removing numbers, punctuation, repeated characters, stopwords , and turn all words into lowercase, and apply lemmatazation. we did 2 types of toknization.

Algorithms

Data manipulation and cleaning.

-remove numbers

-remove punctuation

-remove repeated characters

-turn all words into lowercase

-remove stopwords

-apply lemmatazation

there are two types of toknization, first is called counter Count Vectorizer and Term Frequency–Inverse Document Frequency Vectorizer(it-idf)

Models

Logistic regression, k-nearest neighbors, and random forest classifiers were used before settling on random forest as the model with strongest cross-validation performance.

Model Evaluation and Selection

We split into 55/45 train and test respectively. The training dataset has 20,253 data points and the test dataset has 24,7509 data points after the test/train split. All scores reported below were calculated with 10-fold cross validation on the training portion only.

Algorithm	Accuracy	Precision	Recall	F-1 Score
Logistic Regression cv	0.997	0.997	0.998	0.997
K-Nearest Neighbors cv	0.841	0.899	0.744	0.814
Random Forest cv	0.991	0.990	0.992	0.991
Logistic Regression itidf	0.986	0.984	0.987	0.985
K-Nearest Neighbors itidf	0.908	0.903	0.901	0.902
Random Forest itidf	0.990	0.988	0.991	0.989

ROC Curve

Tools

- Data manipulation and cleaning : Pandas , Numpy.
- Plotting : Seaborn, Plotly and Matplotlib.
- Modeling : sklearn.
- Tokenize : nltk and sklearn.

Communication

In addition to the slides and the visuals included in the presentation, we will submit our code and proposal.