miladshiraniUCB / Fake-Real-News-Classification Public

☆ 0 stars 0 forks

☆ Star ▾ Unwatch ▾

Code Issues Pull requests Actions Projects Wiki Security

main ▾ ···

miladshiraniUCB Data ··· 4 minutes ago 34

View code

README.md

Fake-Real-News-Classification

Advances in technology and social medias have made access to sources of information easier compared to decades ago. In the past, reporters gathered information about an event, then the news was published by a publishing company., By contrast, these days people have the ability to post and publish any news they are exposed to. This has advantages and disadvantages. On one hand, social media companies such as Twitter have made it easier to post and publish news regarding an event much faster than ever; this makes many people more informed. On the other hand, this easy way of publishing and posting news resulted in the existence of an enormous volume of fake news. Therefore, it is important for these platforms to be able to filter out fake news by using different methods.

One main approach to filtering out fake news is using machine learning models such as logistic regression, decision trees, and neural networks. To use these models, it is important to convert text to numerical data. To convert words to numbers, first we need to clean the data and remove the words or information that may not help us categorize the data. After that, we must convert cleaned sentences to individual words called “tokens.” This allows us to design a map that transforms tokenized data into numbers. There are several libraries that we may introduce for these purposes such as NLTK, Spacy, Textbloob etc. In this work, however, we will use NLTK to tokenize the data.

Data Sources

We use 76525 news and tweets which are from different online sources. These sources are:

1. Fake and real news dataset. This source contains 44898 data.
2. Source based Fake News Classification. This source contains 2096 data
3. REAL and FAKE news dataset. This source contains 6335 data.
4. GitHub Repo. This source contains 23196 data.

Project Structure

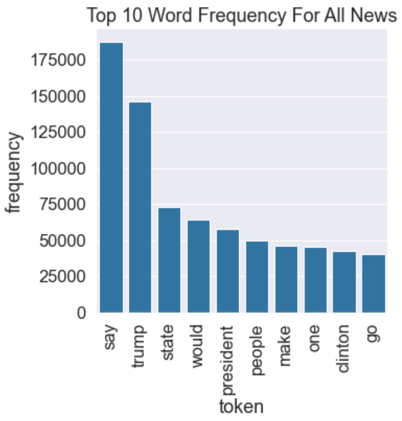
The structure of this project is as follows:

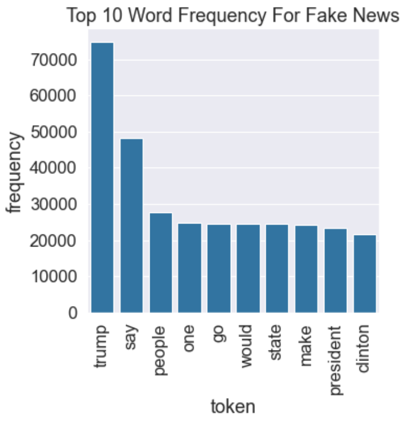
1. **EDA-part-1-Cleaning-Tokenization-Lammatization**. In this notebook we clean the data, and then we tokenize and lemattize the data. This notebook is located in the “EDA” folder.
2. **EDA-part-2-Visualization**. In this notebook, we use the cleaned data and will visualize the top-10 words used in the whole dataset as well as top-1o words in fake news and true news. We also perform some statistical tests to see whether or not some columns of the datasets are coming from a same population. This notebook is located in the “EDA” folder.
3. **Modeling**. In this notebook we first calculate the term frequency–inverse document frequency (TF-IDF) for about 10,000 tokens and then we will train several machine learning models, namely, Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM on training data and we evaluate their performances on the test sets. This notebook is located in the “Modeling” folder.
4. **Modeling_GloVe**.In this notebook, we use a vector representation of words called “Global Vectors for Word Representation” (for short GloVe) prepared by Stanford University to train our machine learning models and then test the performance of the model on testing set. This notebook is located in the “Modeling” folder.
5. **Modeling_NN**. In this notebook, we use TensorFlow to design neural networks to be trained on training data and then we evaluate their performance on testing data. This notebook is located in the “Modeling” folder.

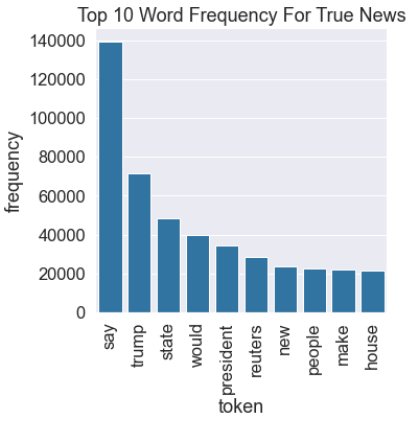
6. **Modeling_NN_Transfer_Learning.** This is the last series of modeling notebooks. In this notebook, we use available embedding layers from TensorFlow Hub and will use them in a neural network to train the rest of the neural network and test its performance on the testing data.

Some Visualizations

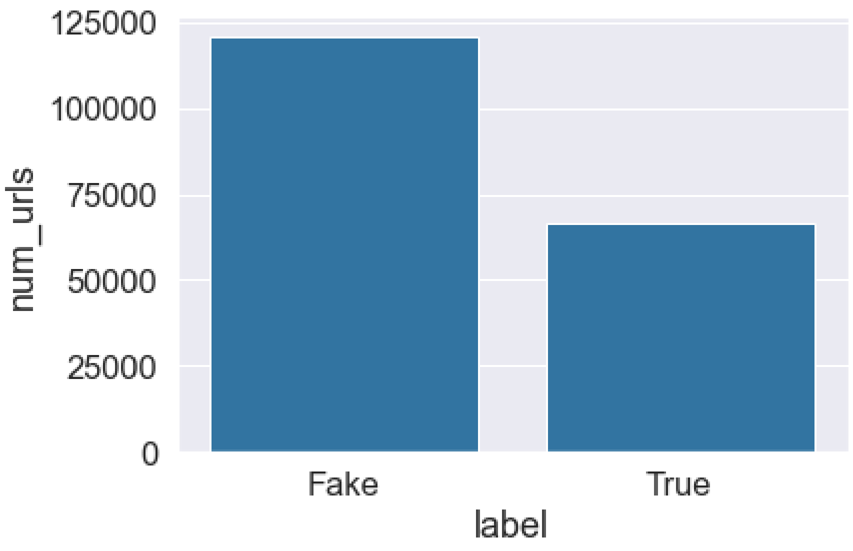
After performing EDA steps, we realized that the top-10 most common words in the news are:



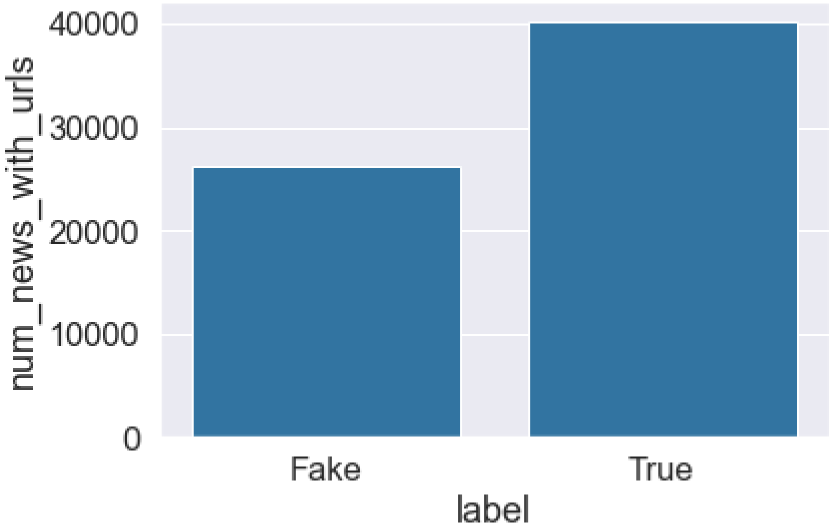




Moreover, we realized that the number of url links in the fake news are more than the total number of url links in the real news as shown below:



However, we see that the total number of real news that contain url links are more than the total number of fake news with url links as shown below:



Final Model

The model we introduce as a final model is the **3rd model** from the last notebook (Modeling_NN_Transfer_Learning) and the result of the model on the training and test sets are shown below.

The models introduced in this project may be used by social medias such as twitter to detect and filter out fake news from real news.

Recommendations to Improve the Model

We would recommend to gather more data and also to use combination of different layers in addition to the embedding layer such as LSTM or Conv1D and also we would recommend to optimize the hyperparameters of the model by using optuna .

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- ● **Jupyter Notebook** 100.0%