DLI Accelerated Data Science Teaching Kit

Lecture 24.1 - Introduction to Team Project

Fake News Detection on Social Media

Social media (e.g., Twitter and Facebook) has become a new ecosystem for spreading news.

Nowadays, people are relying more on social media services rather than traditional media because of its advantages such as social awareness, global connectivity, and real-time sharing of digital information.

Unfortunately, social media is full of fake news. Fake news consists of information that is intentionally and verifiably false to mislead readers, which is motivated by chasing personal or organizational profits.

Fake news detection is to determine the truthfulness of the news by analyzing the news contents and related information such as propagation patterns.

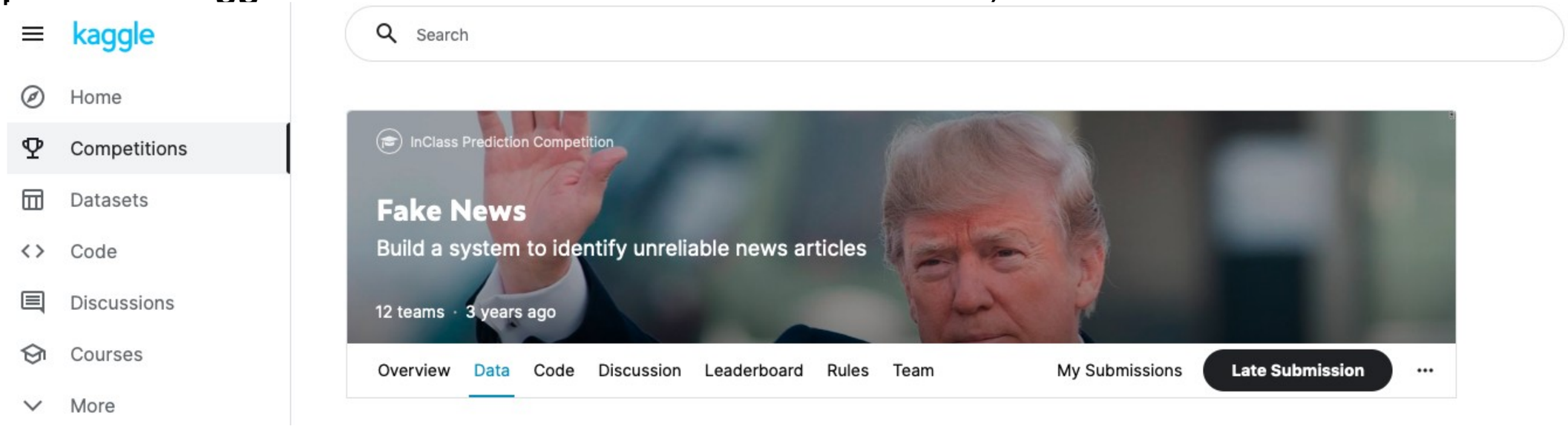
Source: https://abcnews.go.com/US/ways-spot-disinformation-social-media-feeds/story?id=67784438

Fake News Detection via Machine Learning

Treat fake news detection as text classification

- Two classes: Fake and True
- Short texts from social media

Dataset (**https://www.kaggle.com/c/fake-news/data?select=train.csv**)



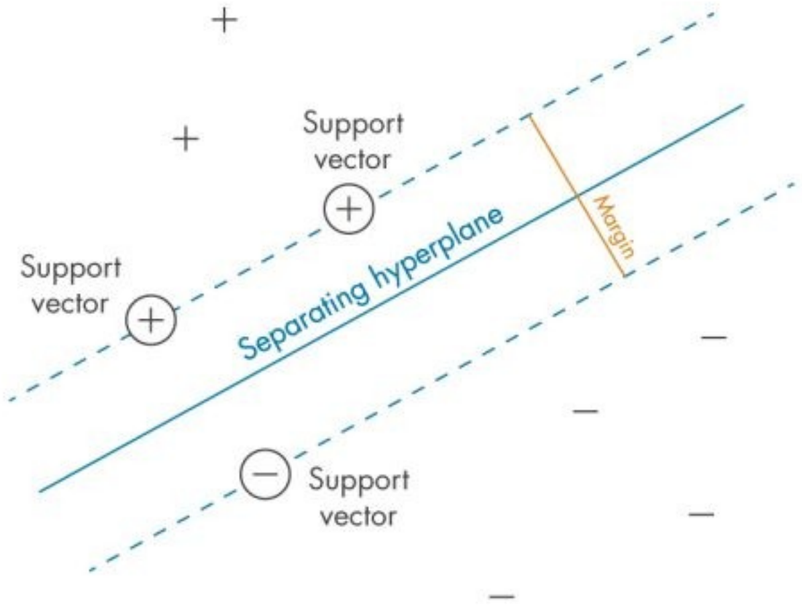
Source: https://www.kaggle.com/c/fake-news/data?select=train.csv

Machine Learning Models for Classification

Support Vector Machine

Random Forest

Logistic Regression



Source: https://www.mathworks.com/discovery/support-vector-machine.html

Data Preprocessing

Five Steps:

- Remove rows with missing values
- Stemming (https://en.wikipedia.org/wiki/Stemming)
- Remove stop words (https://en.wikipedia.org/wiki/Stop_word)
- Extract features with Term Frequency — Inverse Document Frequency (TFIDF) (https://en.wikipedia.org/wiki/Tf–idf) to build samples
- Split the samples into training and testing data with the ratio 0.2

Building Fake News Detection Models

Nine Steps:

- Load training data and testing data
- Train a classifier 1 with traditional random forest from sklearn on the training data and record the training time
- Train a classifier 2 with traditional Logistic Regression from sklearn on the training data and record the training time
- Train a classifier 3 with traditional SVM from sklearn on the training data and record the training time
- Train a classifier 4 with random forest from cuML on the training data and record the training time

Building Fake News Detection Models

Nine Steps:

- Train a classifier 5 with Logistic Regression from cuML on the training data and record the training time
- Train a classifier 6 with SVM from cuML on the training data and record the training time
- Complete the inference on testing data with these six models and record the classification performance such as accuracy
- Compare the training time and the classification performance

DLI Accelerated Data Science Teaching Kit

Thank You