Many datasets are available online related to false news (as discussed in the examples above) and on Kaggle. For building a fake news detection machine learning project, we will use the false news competition dataset on Kaggle.

Image for Dataset For Detecting False News Using Data Science

This CSV file contains the title and text of the news article, the author's name, and a label that identifies certain ideas in the article as reliable or unreliable. Most misinformation-related datasets will have a similar labeling and tabular structure- however, there are datasets with multiple labels that signify different levels of reliability (from totally reliable to totally unreliable).

The first step is to import the basic requirements in Python. We will use Pandas to load the dataset.

Dataset Description

train.csv: A full training dataset with the following attributes:

- **id**: unique id for a news article
- **title**: the title of a news article
- **author**: author of the news article
- **text**: the text of the article; could be incomplete
- **label**: a label that marks the article as potentially unreliable
  - 1: unreliable
  - 0: reliable

Image for Importing Dataset For Detecting False News Using Python

Since the training and testing set only has the target label defined, we will split the data from the training and testing set into a train and validation test set.

Image for Reading Training Dataset For Detecting False News

As expected from the description, a quick look at the data shows what we must expect. Let's look at some basic statistics before we move forward with modeling.

Unbalanced classes, including fake news, are a challenging problem in anomaly detection tasks. Fortunately, the difference in the frequency of labels for this dataset is not too problematic. We can use the training set as it is. Different techniques one can apply in other cases include SMOTE for oversampling, random undersampling, or weighted evaluation during training, which gives higher weights to underrepresented classes during loss calculation.

Next, we see how the length of the articles (from the word count) is distributed in the dataset. The histograms below show that most articles are under 1000 words long, with an average of 804 words (calculated separately).

In traditional news making procedures, very limited and authorized individuals are involved and newspapers, radio, television were the only source of news. Due to these reasons news, credibility and authenticity are preserved. But in the era of internet, social network is becoming a news source of news. Easy and free access to these social networks makes the task of fabricating fake news and manipulating news a very effortless task. There is no authorize control point of these manipulated fake news which creates a question over there credibility and authenticity. The ease of getting direct news from the platform they mostly use has attracted the user. The reason to spread fake news can be social, political, and economical. Fake news in business can affect the stocks of the company leading to a huge capital loss. During the election campaign, fake news is used as a weapon against each other in a political war to defame the opposition. The most adverse effect is seen when it is used to spread communal hates which leads to riots. The Delhi riots are the best example of the destruction caused by fake news.

Fake news about the COVID-19 in India lead to an attack on the medical team in various parts of the country and thus making the fight against the virus weak. The rate at which it spread is very fast due to which controlling the spread manually is not possible. There is no platform via which the user can check the credibility and authenticity of the news and where authorities can directly inform about the fake news prevailing. Due to which people can believe in the news which can be a trouble for them and as well for society also. In the existing system, the action is taken after the adverse impact had already hit society. The proposed platform is useful for both common people and official authorities to prevent the spread of rumours in form of news.