Naive Bayes classifier for Fake News recognition

Fake news are defined by the New York Times as "a made-up story with an intention to deceive", with the intent to confuse or deceive people. They are everywhere in our daily life, and come especially from social media platforms and applications in the online world. Being able to distinguish fake contents form real news is today one of the most serious challenges facing the news industry.

Naive Bayes classifiers [1] are powerful algorithms that are used for text data analysis and are connected to classification tasks of text in multiple classes.

The goal of the project is to implement a Multinomial Naive Bayes classifier in R and test its performances in the classification of social media posts. The suggested data set is available on Kaggle [2].

Possible suggested lables for classifying the text are the following:

- True - 5

- Not-Known - 4

- Mostly-True - 3

- Half-True - 2

- False - 1

- Barely-True - 0

The Kaggle dataset [2] consists of a training set wth 10,240 instances and a test set wth 1,267 instances.

Bibliography

[1] C. D. Manning, Chapter 13, *Text Classification and Naive Bayes*, in *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[2] Fake News Content Detection, KAGGLE data set: `https://www.kaggle.com/datasets/anmolkumar/fake-news-content-detection?select=train.csv`