Report nr.1

Repo: https://github.com/martentyrk/FakeNewsDetection
Group ID: D15

Business Understanding

Identifying your business goals:

- Background: Fake news is false or misleading information presented as news. These fake news articles tend to come from satirical news websites or individual websites with an incentive to propagate false information, damaging the reputation of a person or entity either as clickbait or to serve a purpose, or making money through advertising revenue.
- Business goals: The purpose of the work is to come up with a fake news detector that can be utilized by users to detect if a random article is fake news or real news.
- Business success criteria: Our detector is able to distinguish between real and fake news with an accuracy of over 93%.

Assessing your situation:

- Inventory of resources: ~18000 Fake news articles and ~20000 Real news articles. 5 laptops. 1 desktop computer 3 smartphones. 3 semi-hardworking university students. 3 lab instructors and 1 lecturer for technical support. Any software that's open source.
- Requirements, assumptions, and constraints: End date for completion December 17. May have to refer to the author's of the articles to be able to use them. As an end goal, create a model that is able to distinguish between real and fake news with an accuracy of over 93%.
- Risks and contingencies: Loss of access to dataset - Use new one from https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset

Loss of internet/electricity - Attempt to move in with another project member, if prevented by corona wait out the contingency and if needed postpone project deadline.

- Terminology

Bootstrapping - Training data sets are created by re-sampling with replacement from the original training set, so data records may occur more than once.

Confidence - Confidence of rule “B given A” is a measure of how much more likely it is that B occurs when A has occurred.

degree of fit - A measure of how closely the model fits the training data. A common measure is r-square.

dependent variable - The dependent variables (outputs or responses) of a model are the variables predicted by the equation or rules of the model using the independent variables (inputs or predictors).

Deployment - After the model is trained and validated, it is used to analyze new data and make predictions. This use of the model is called deployment.

external data - Data not collected by the organization, such as data available from a reference book, a government source or a proprietary database.

internal data - Data collected by an organization such as operating and customer data.

Interaction - Two independent variables interact when changes in the value of one change the effect on the dependent variable of the other.

Mean - The arithmetic average value of a collection of numeric data.

Median - The value in the middle of a collection of ordered data. In other words, the value with the same number of items above and below it.

Mode - The most common value in a data set. If more than one value occurs the same number of times, the data is multi-modal.

Noise - The difference between a model and its predictions. Sometimes data is referred to as noisy when it contains errors such as many missing or incorrect values

Overfitting - A tendency of some modeling techniques to assign importance to random variations in the data by declaring them important patterns.

- Costs and benefits: 0 €,

Defining your data-mining goals:

- Data-mining goals: We use datasets that we found from Kaggle.
- Data-mining success criteria: We selected datasets with usability higher than 8.5.