How to solve it

- **Classification problem**
  - News Report (***document***) → ***Class***: [FAKE, REAL]
- **Try text-related classifiers**
  - Naive Bayes
  - MaxEnt
  - SVM
- **NLTK+SKLearn provides you anything you need**
  - NLP Pre-processing
  - Classifiers
  - N-grams

Dataset

- **fake_or_real_news_training:**
  - **ID**: ID of the news
  - **Title**: Title of the news report
  - **Text**: Textual content of the news report
  - **Label**: Target Variable [FAKE, REAL]
  - **X1, X2:** additional fields
- **fake_or_real_news_test:**
  - **ID, title and text**
  - **Predict Label**

Advices

- **Take a look to the data**
  - Check your data loading process
  - News have 2 levels of text (title and text)
- Try the **pre-processing methodologies** we have **in class**
- **TF-IDF** seems to be better (but try it!)
- **N-grams** pay the effort
- Less than 90-92%? **Try again**

Advices/Warnings

- Avoid ML mistakes
- Explain anything you do
- Try different approaches and compare results
  - Classifiers
  - NLP Pipelines
- Analyze your results

- **Submission** (Send **everything** please)**:**
  - **CSV with your predictions**
    - **News_id (ID), prediction[FAKE, REAL]**
  - **Notebook**
- Send me something that **actually works**
- **Grading:** 50% results – 50% notebook