Dataset and Objectives - Machine Learning Project 2020

Harry Chirayil, Christopher Keim, Stefan Schmutz

Dataset

The dataset includes **news headlines** (title and short teaser) from two online news sites 20min.ch and nzz.ch.

Format

date_time	date and time of web scraping in Coordinated Universal Time (UTC)
object	type of text, either "title" or "teaser"
order	order headline appeared on website, makes it possible to link title and teaser
text	full text, all in lowercase and punctuation marks removed

Following are the first few rows of the `headlines_nzz.csv` raw data:

```
## # A tibble: 65,322 x 4
##   date_time           object order text
##   <dttm>              <chr>  <dbl> <chr>
## 1 2020-01-21 05:00:14 title      1 die konzernlenker sind so pessimistisch wie ~
## 2 2020-01-21 05:00:14 title      2 wef das wohl wichtigste treffen für die schw~
## 3 2020-01-21 05:00:14 title      3 die absperrungen sind aufgestellt die geschä~
## 4 2020-01-21 05:00:14 title      4 präzisionsschützen und ein sprengstoffkomman~
## 5 2020-01-21 05:00:14 title      5 die luanda leaks bieten einblicke in ein kap~
## # ... with 65,317 more rows
```

Data collection

All titles and teasers (if available) from the titlepage were collected twice a day (6am and 6pm local time) between 2019-02-04 and 2020-01-21.
While the amount of unique titles are comparable between the two sources, there were more teasers available from the 20min titlepage (see Table 1).

Table 1: Sample size per class

source	object	n_unique
nzz	Title	26'656
20min	Title	27'874
nzz	Teaser	5'484
20min	Teaser	27'295

Objective

Can we predict the source of a title or teaser based on the chosen words?
We might want to try title or teaser separately, or a combination and compare the performances.

Data visualisation

To gain some insights, basic data visualisation was done on the dataset where replicates (stories listed multiple times) were only kept once.
Figure 1 shows the headline length (words) distributions within the different categories. It can be seen that nzz headlines are typically a bit longer and with a larger spread compared to 20min headlines.
Figure 2 compares the word frequencies from the two sources. Stop words, words that are very common and typically not very useful for classification were removed (source: github.com/solariz/german_stopwords).
It's visible that many of the most frequently used words occur at different rates in headlines of the two news-sites. This could hint that it's possible to classify headlines based on the words used.

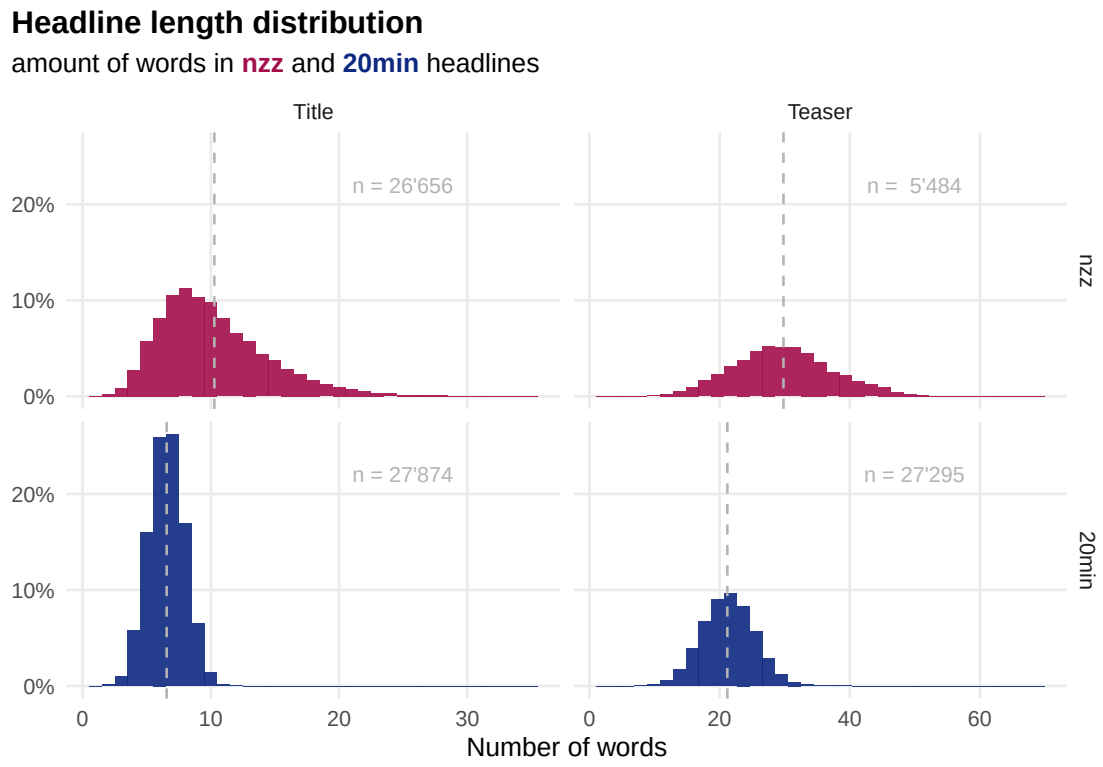

Figure 1: Length distribution of headlines in words. Dotted lines represent means per group.

Heads up

Some headlines might be advertisement. We could detect those by looking at how often they were visible. We'd expect advertisement to appear more frequently compared to actual news headlines.

For the 20min dataset we have a teaser for almost every title, this is not the case for the nzz dataset where very often we only have a title without teaser.

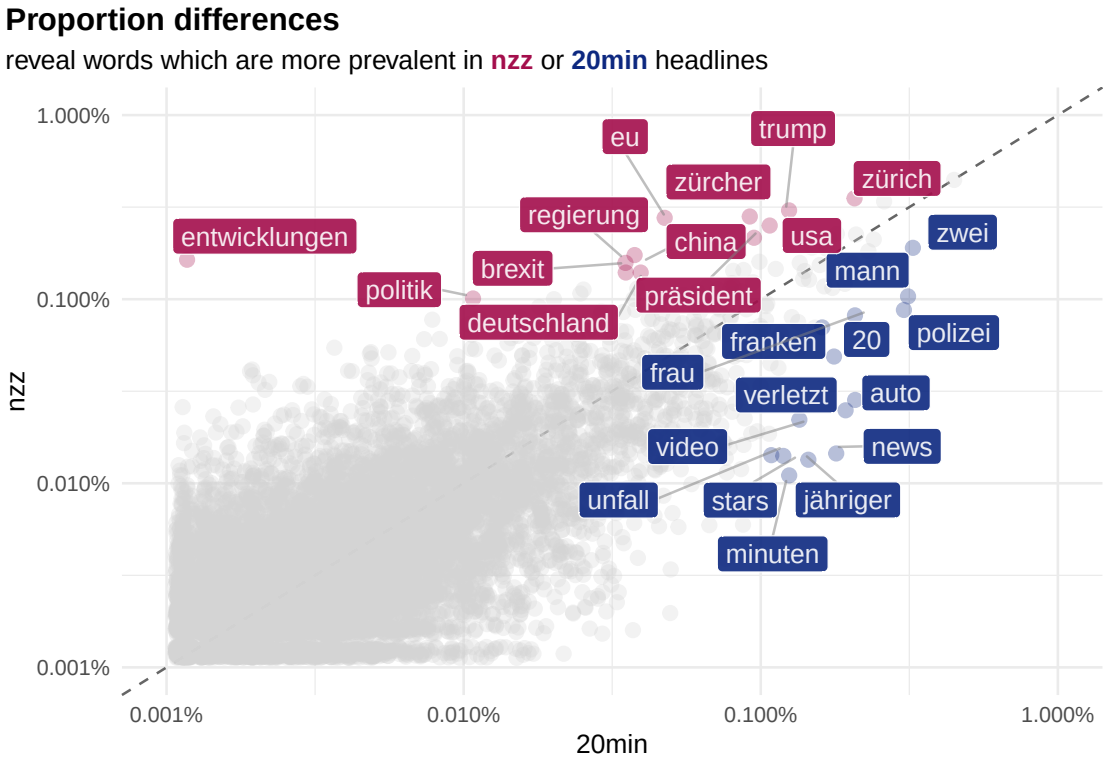

Figure 2: Word frequencies in title and teaser per source (not including stop words)

Chosen Estimator

Following the flow diagram below (see Figure 3), we might want to try a Linear Support Vector Classification or Naive Bayes.

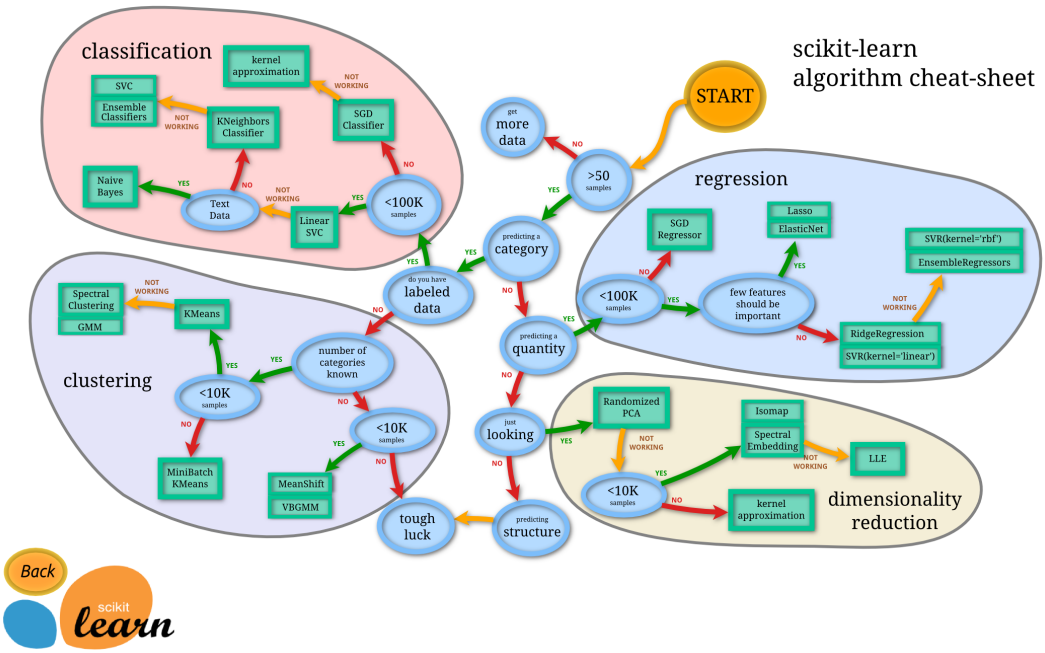

Figure 3: Choosing the right ML estimator, source: scikit-learn.org