Data Challenge Workshop: Classification of business-related News Articles

Christian Ritter, Data Science Accelerator, Enterprise Statistics Division

www.statcan.gc.ca

August 20, 2018

STATISTICS CANADA

ONE HUNDRED YEARS AND COUNTING

STATISTIQUE CANADA

CENT ANS BIEN COMPTÉS

christian.ritter@canada.ca

Code & data available at
https://github.com/chritter/Talks/tree/master/2018/Data_Challenge_UOttawa

Statistics Canada Statistique Canada

Data Science Accelerator

Enabling you.

Statistique Canada Statistics Canada

Canada

II. Our team & mandate

Joanne Yoon, BSc

Monica Piccard, MSc
Unit Head

Stan Hatko, MSc

Sevgui Erman, PhD,
Assistant Director,
DSA Lead

Christian Ritter, PhD

Saeid Molladavoudi, PhD

Nic Denis, MSc

Motivation

The problem

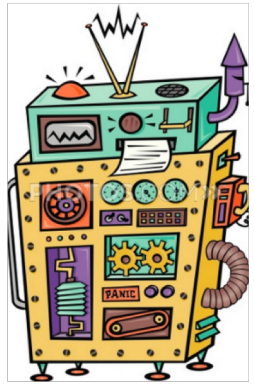
- Business analysts in enterprise statistics division spend large amounts of time to search for information
- Amount of news exponentially growing
- What happened? When? Where? Who?

The solution

- Automatic analysis of text
- Capture more information faster
- Event and trend detection of unexpected news
- Knowledge extraction for companies and industries

Approach

Event detector



Filter non-economic relevant news
(keywords, publisher)

Deduplication & grouping

Canadian News Sources

Event type
Companies involved

Merger & Acquisition

Bankruptcy

events

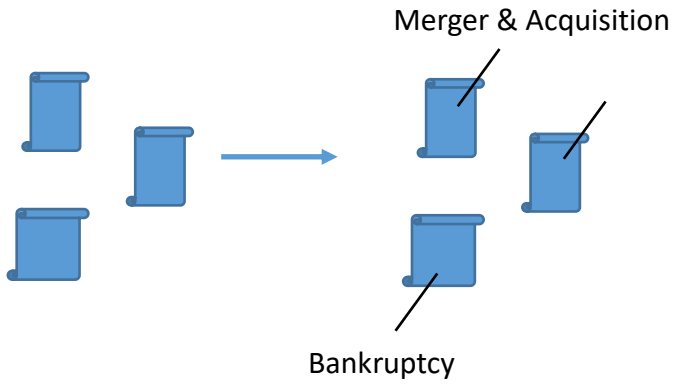
Company X
Merger
1/1/11

time

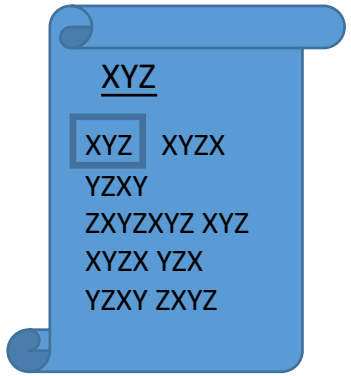
The event detector: Supervised machine learning

A) Creating a labeled dataset

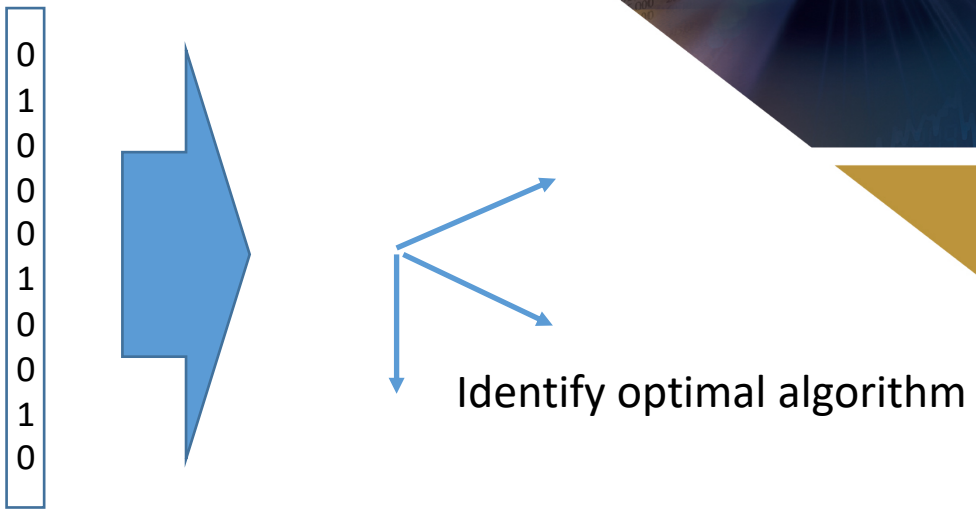
News Articles



B) Feature engineering



C) Classification



- Document-level event detection
- Name-entity recognition for deduplication and to identify companies involved

Feature engineering

- Major challenge in data science workflow
- Design characteristics from which your algorithm can learn to distinguish between classes
- In our case: Lexical and syntactic features
- For example:
  - **Type of word, compound (n-gram) words**
  - **Weight occurrence of each word: Term frequency–Inverse document frequency (tfidf)**

t: term
d: document
D: all documents

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Linear Regression

- Advantage to start with simple model
- Easy to interpret due to linear relationship with input x and output y
- Fast baseline model
- Logistic regression for classification

Decision threshold

Feature 0

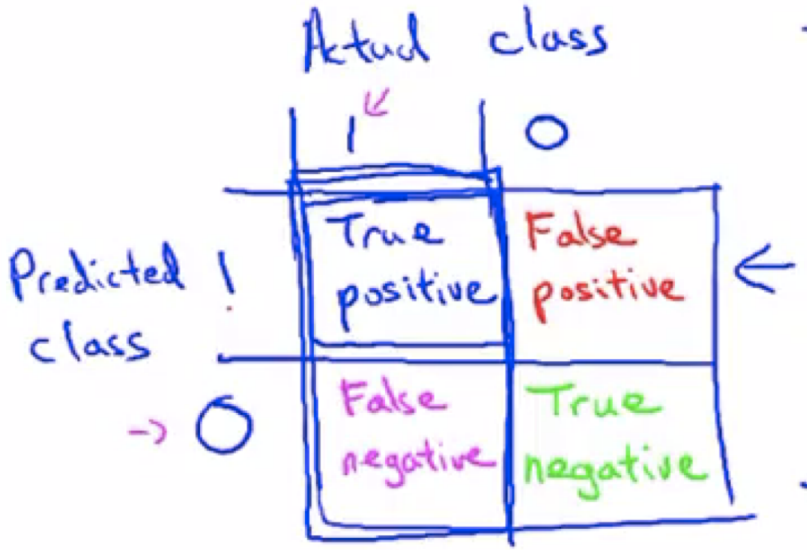
Weight 0

$$\hat{y} = w[0]x[0] + w[1]x[1] + \ldots + w[p] * x[p] + b > 0$$

Evaluation Metrics

Precision/Recall

$y = 1$ in presence of rare class that we want to detect



→ **Precision**

(Of all patients where we predicted $y = 1$, what fraction actually has cancer?)

$$\frac{\text{True positives}}{\#\text{predicted positive}} = \frac{\text{True positive}}{\text{True pos} + \text{False pos}}$$

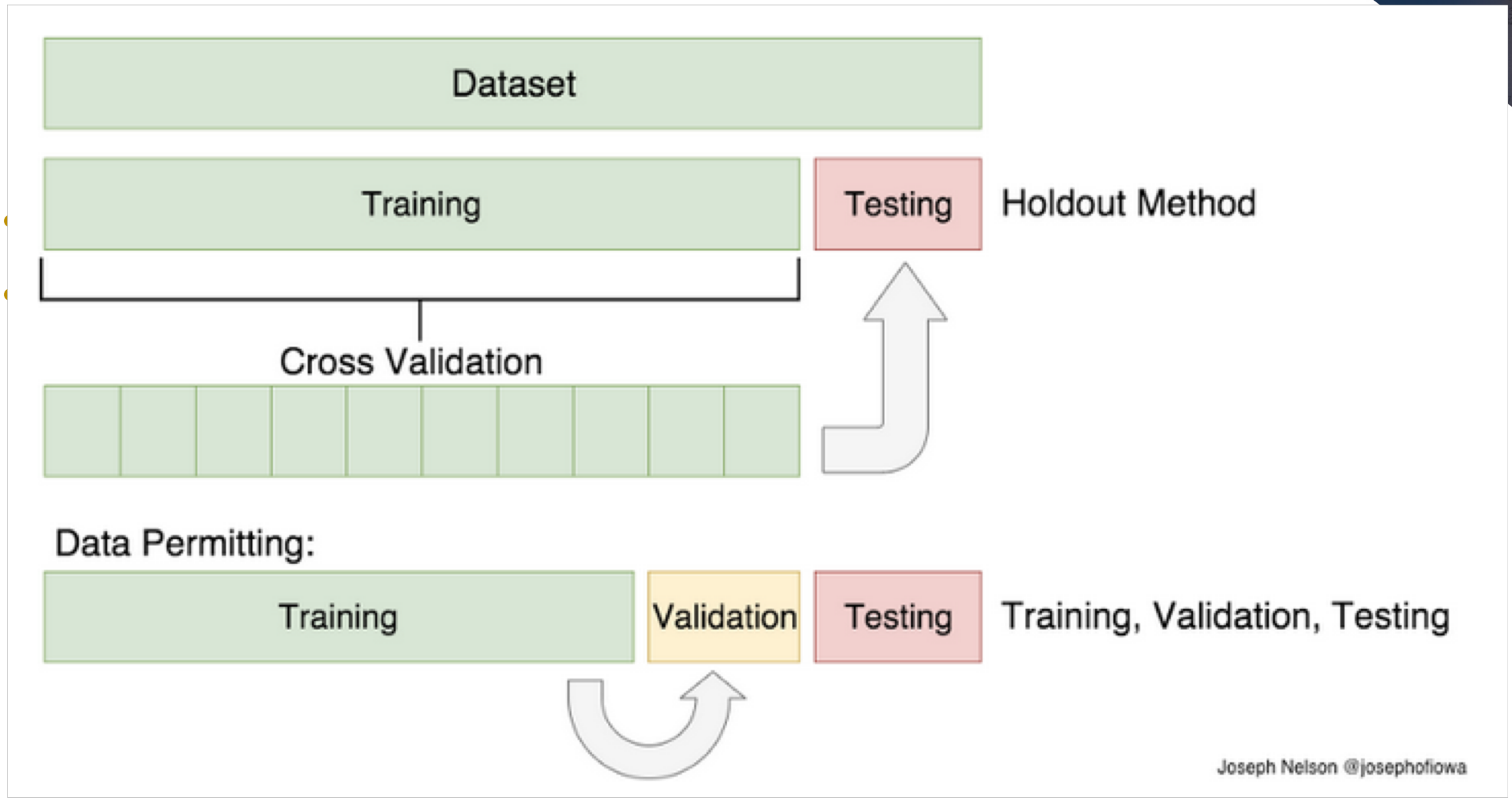
→ **Recall**

(Of all patients that actually have cancer, what fraction did we correctly detect as having cancer?)

$$\frac{\text{True positives}}{\#\text{actual positives}} = \frac{\text{True positives}}{\text{True pos} + \text{False neg}}$$

Slide by Andrew Ng (Machine Learning on Coursera)

Cross validation



Implementation tips

- Do you have a regression or classification problem?
- Do you have training data and what is the quality?
- Could you leverage open data and algorithms?
- Which features could help to distinguish between the classes? Features often matter more than algorithms
- Each model needs cross validation

Q&A

- Code & data available at https://github.com/chritter/Talks/tree/master/2018/Data_Challenge_UOttawa
- Could you adopt classification in your problem case?