Text and Tax

Tobias Elvermann[1], Carina Hausladen[2], Agajan Torayev[1]

University of Bonn[1], University of Cologne[2]

June 12, 2018

introduction

- cooperation with the chair for Behavioral Accounting, Taxation and Finance of the University of Cologne
- tax risk
- tax risk management

data

- annual reports of the last ten years from 600 companies listed in the STOXX Europe 600
- keyword search for "tax", "risk"
- resulted in 27'000 rows
- three labels (manually assigned) $\rightarrow$ gold standard data!

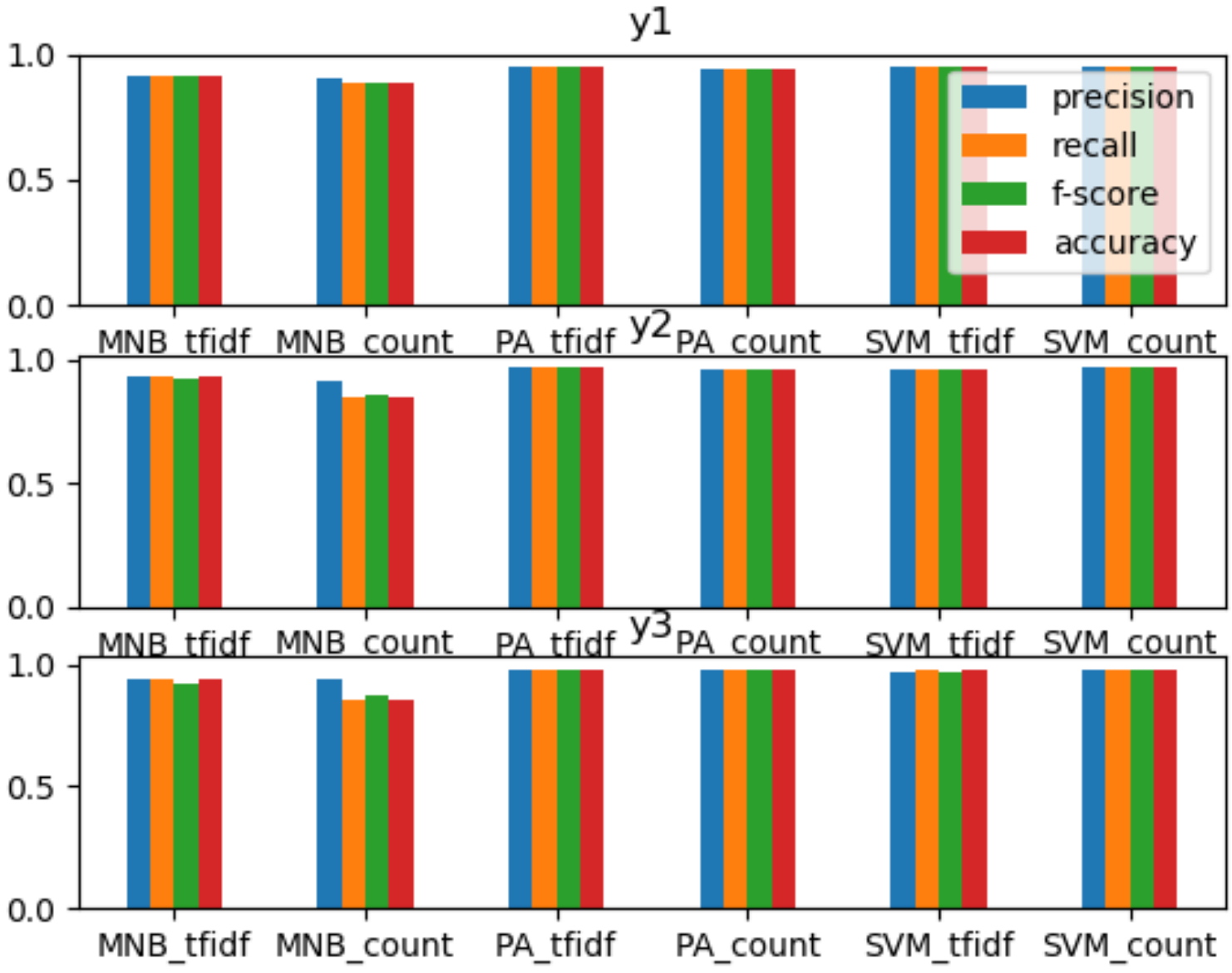
goal

- train an algorithm to learn the labels $\rightarrow$ introspection of an even more extensive dataset is possible
- redo classification with full annual reports instead of paragraphs
- text summarization of measures on how tax risk management was implemented by companies

methods

- baseline: multinomial bayes (MNB), passive aggressive classifier (PA), support vector machine (SVM)
- advanced model: CNN + LSTM
- method of comparison: precision, recall, f-score, accuracy and confusion matrix

baseline models



baseline models

Confusion matrices for the best performing models for each of the three labels were selected.

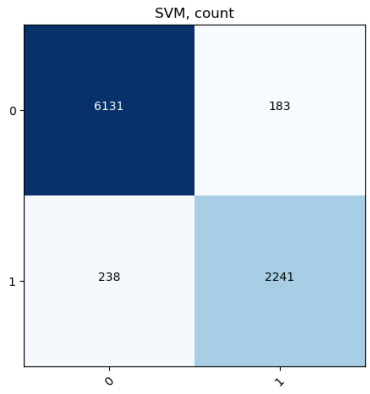


Figure 1: best model for y1: SVM + count

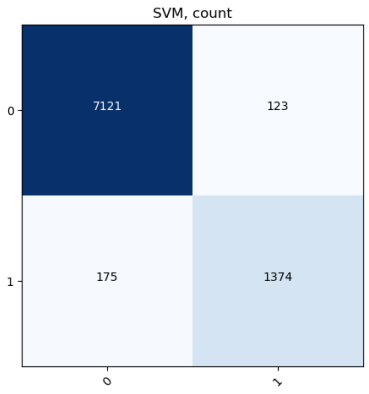


Figure 2: best model for y2: SVM + count

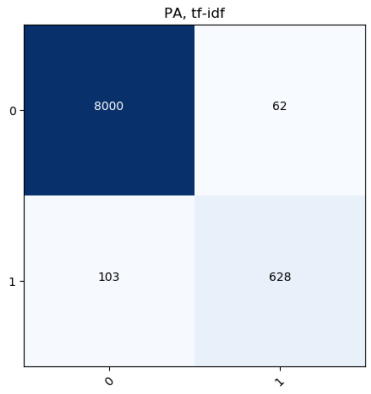


Figure 3: best model for y3: PA + tfidf

comparing features: tfidf

y1	y2	y3
tax1	tax1	tax1
risk	risk	group
risks	group	risk
group	risks	risks
x92	x92	management
financial	changes	x92
management	management	financial
changes	financial	compliance
business	authorities	department
legal	business	committee
company	laws	business
authorities	law	control
x952013	legal	x952013
taxes3	taxdiscountrateto	policy
related	company	audit
law	taxes3	authorities
compliance	subject	legal
laws	regulations	internal
control	compliance	changes
regulations	committee	policies

comparing features: count

y1	y2	y3
tax	tax	tax
risk	risk	group
risks	group	risk
group	risks	risks
x92s	x92s	management
financial	changes	x92s
management	management	financial
changes	financial	compliance
business	authorities	department
legal	business	committee
company	laws	business
authorities	law	control
x95	legal	x95
taxes	taxation	policy
related	company	audit
law	taxes	authorities
compliance	subject	legal
laws	regulations	internal
control	compliance	changes
regulations	committee	policies

CNN + LSTM

- Keras (Tensorflow backend)
- 1 CNN + 1 LSTM + 1 FCC
- For 27K data: about 90 seconds in CPU

confusion matrix

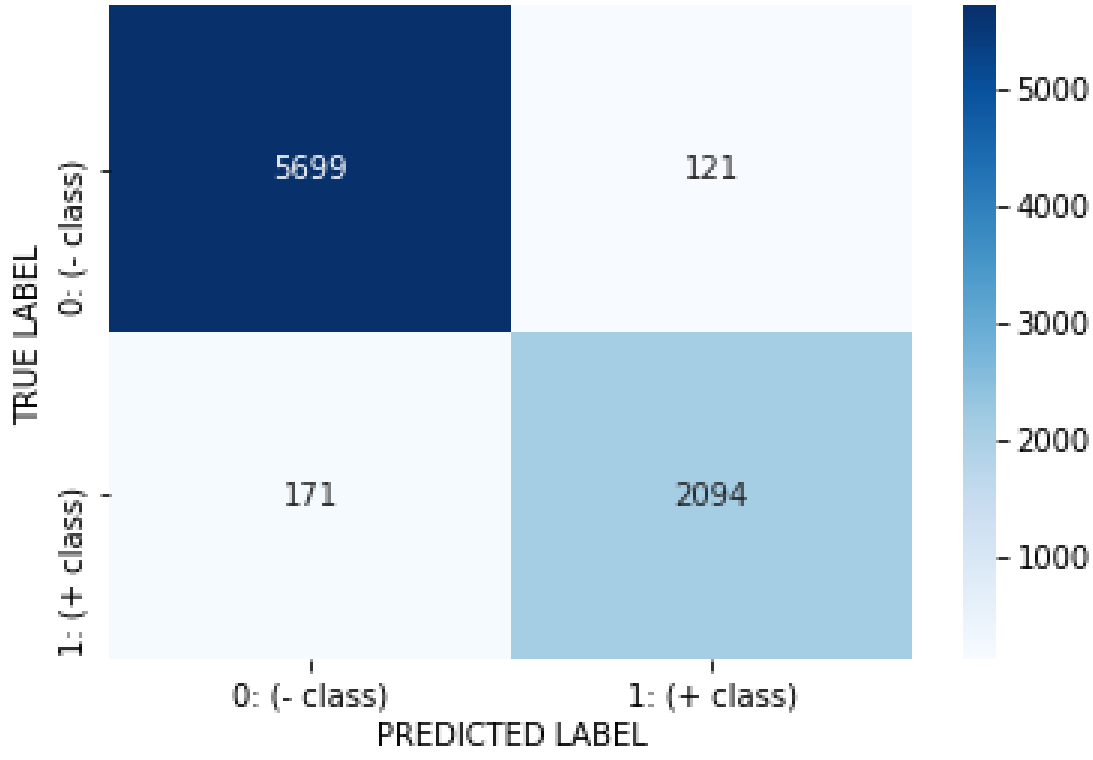


Figure 4: confusion matrix CNN+RNN for y1