Exploiting the Web Presence of Enterprises to Improve NACE Code Classification

Johannes Gussenbauer
Johannes.Gussenbauer@statistik.gv.at

Alexander Kowarik
Alexander.Kowarik@statistik.gv.at

WIN 2025 CONFERENCE Danzig, 05.02.2025

www.statistik.at

Web Intelligence Network

STATISTIK AUSTRIA
Die Informationsmanager

Outline

- Aim of classification task
- Data acquisition and processing
- Modelling and performance evaluation
- Hierarchical performance measures

Aim of classification task

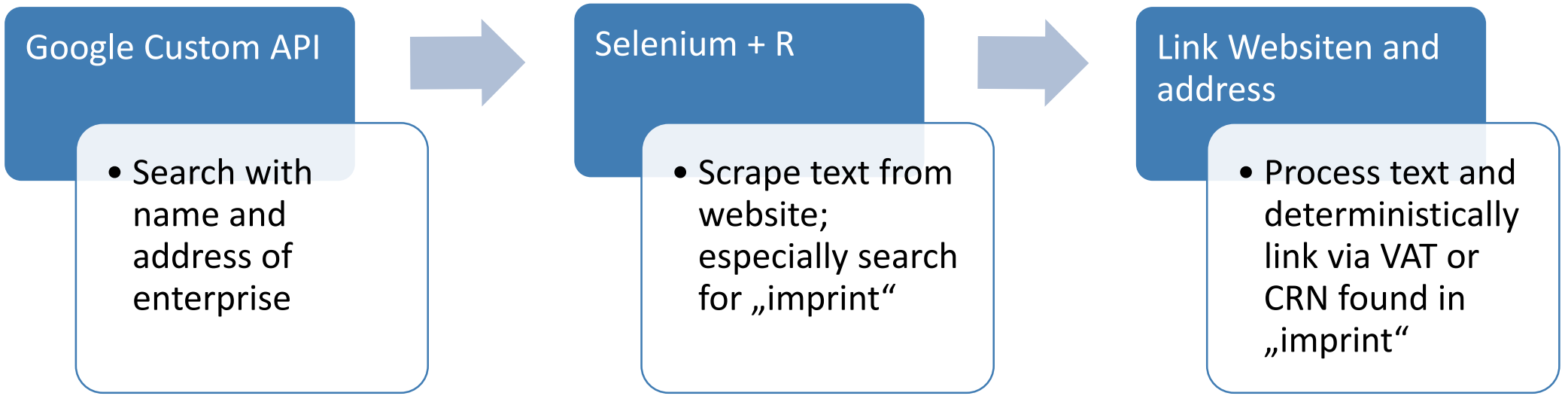
Aim of classification task

- NACE editing labour intensive task + NACE revision coming 2025
- Possible to predict NACE of entrprise using text from enterprise website?
- Test NACE predicion during ESSNet Web Intelligence Network
- Main focus on developing model used in recommendation system for editing task ➔ reduce editing time

Data Acquisition and pre-processing

Data Acquisition

- Collect web data during ICT-survey cycles
  - Collected data from 2019 to 2023 (results limited up to 2021)



STATISTICS | PUBLICATIONS & SERVICES | CLASSIFICATIONS | SURVEYS | DOCUMENTATIONS | PRESS | ABOUT US | INDEX

Website information

Media owner

STATISTICS AUSTRIA
Federal Institution under Public Law
Guglgasse 13
A-1110 Vienna
Tel.: +43 (1) 71128 0
Fax: +43 (1) 71128 7728
office@statistik.gv.at

Company register: FN 191155k, registry court: Vienna Commercial Court
Registered office: Vienna, place of jurisdiction: Vienna
VAT ID No.: ATU37869909

Data Protection Information:

- www.statistik.at

dsgvo@statistik.gv.at

Disclosure in accordance with § 25 Austrian Media Act

Text data processing

- Process collected text from website
- Transform each word with the German morphological lexicon available on https://www.openthesaurus.de/about/download
  - Lemmetization and stemming did not improve classification performance
- Removing all digits and punctuations
- Remove characters not part of the German dictionary
- Remove German stop words.

Modelling & Results

NACE Classification

- Make NACE level 2 prediction using text as features

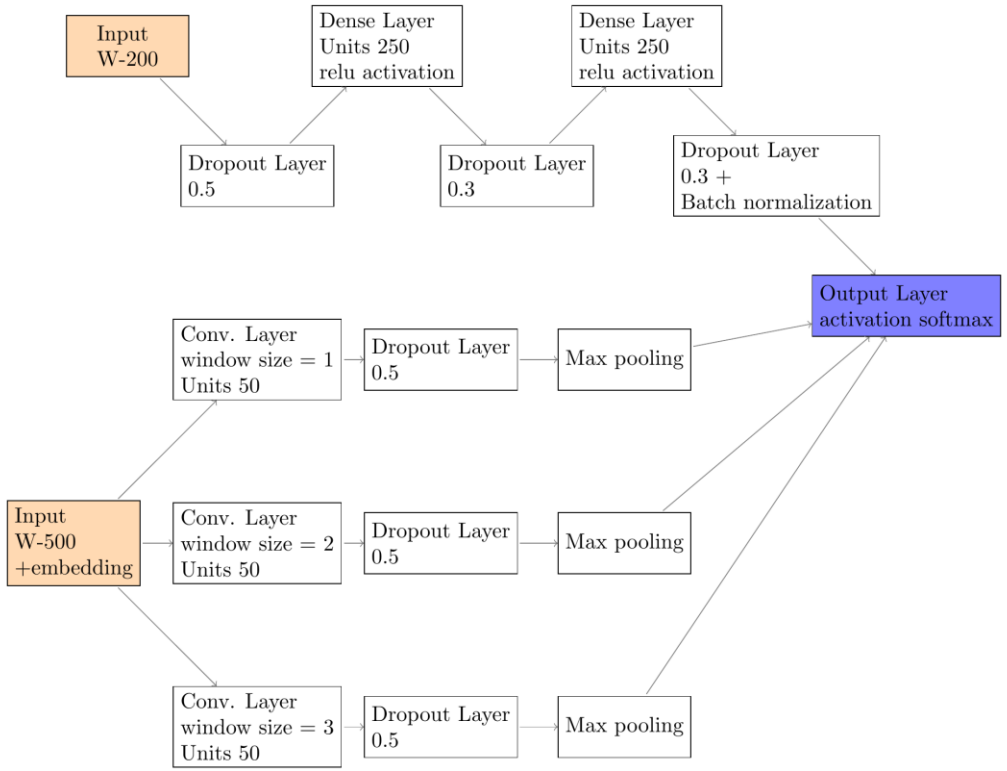
$$NACE = f(Text\ from\ Website)$$

- Pre-processed text contains over 2 Mio different words
  - Use internal dictionary decribing NACE codes?
  - Use feature selection by Uysal (2016)

- Combine global and local feature selection score to determine set of words used for prediction
  - Information Gain (IG); Gini Index (GI); Distinguishing Feature Selector (DFS)
  - Create set of 200 and 500 „important“ words per NACE 2 level code

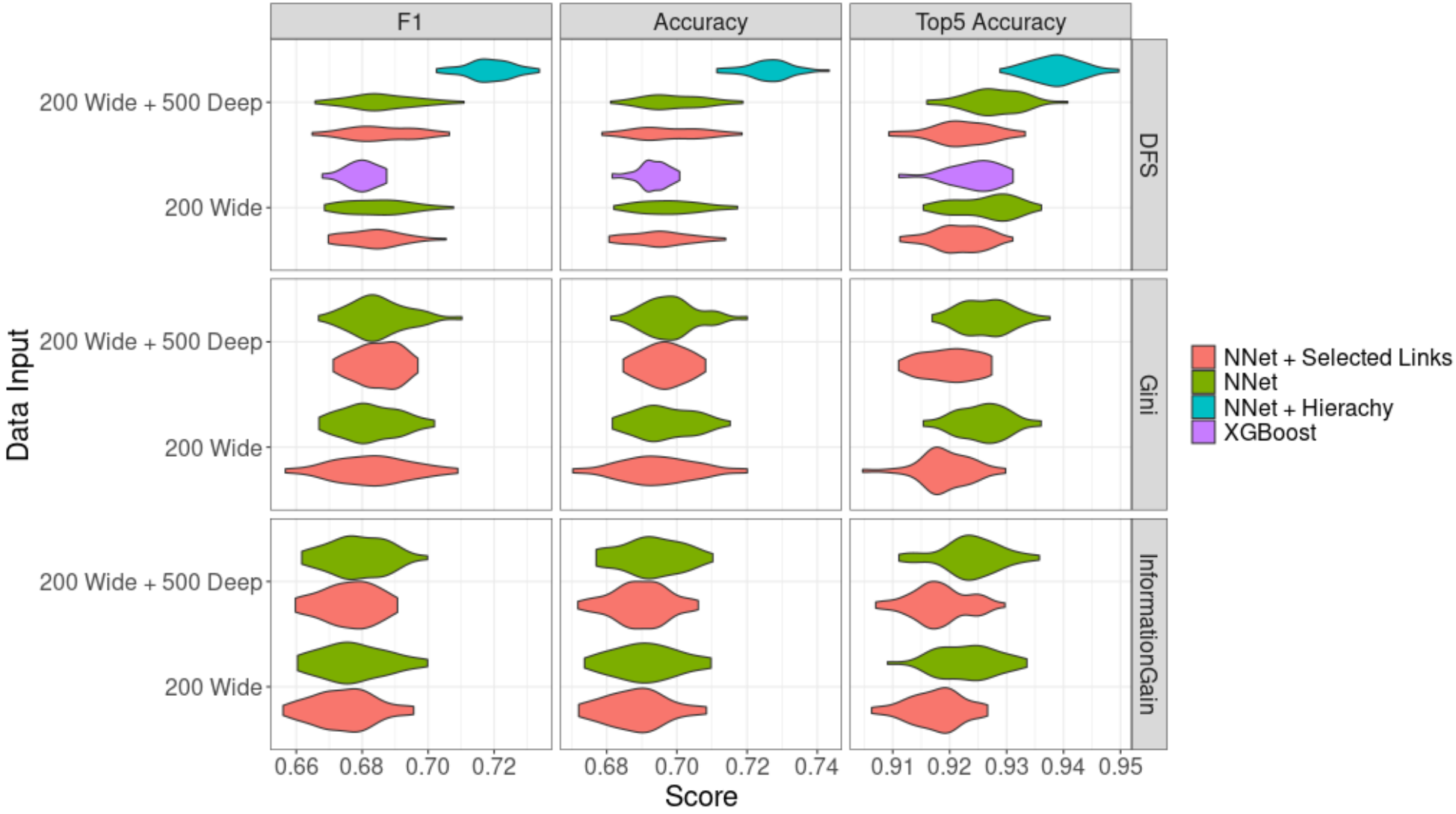
Choice of models

- Neural Network (R-packages keras, tensorflow)
  - Prominently used for NLP tasks
  - Possibility to use so called word embeddings
  - Use combination from One-Hot-Encoded data and word embeddings

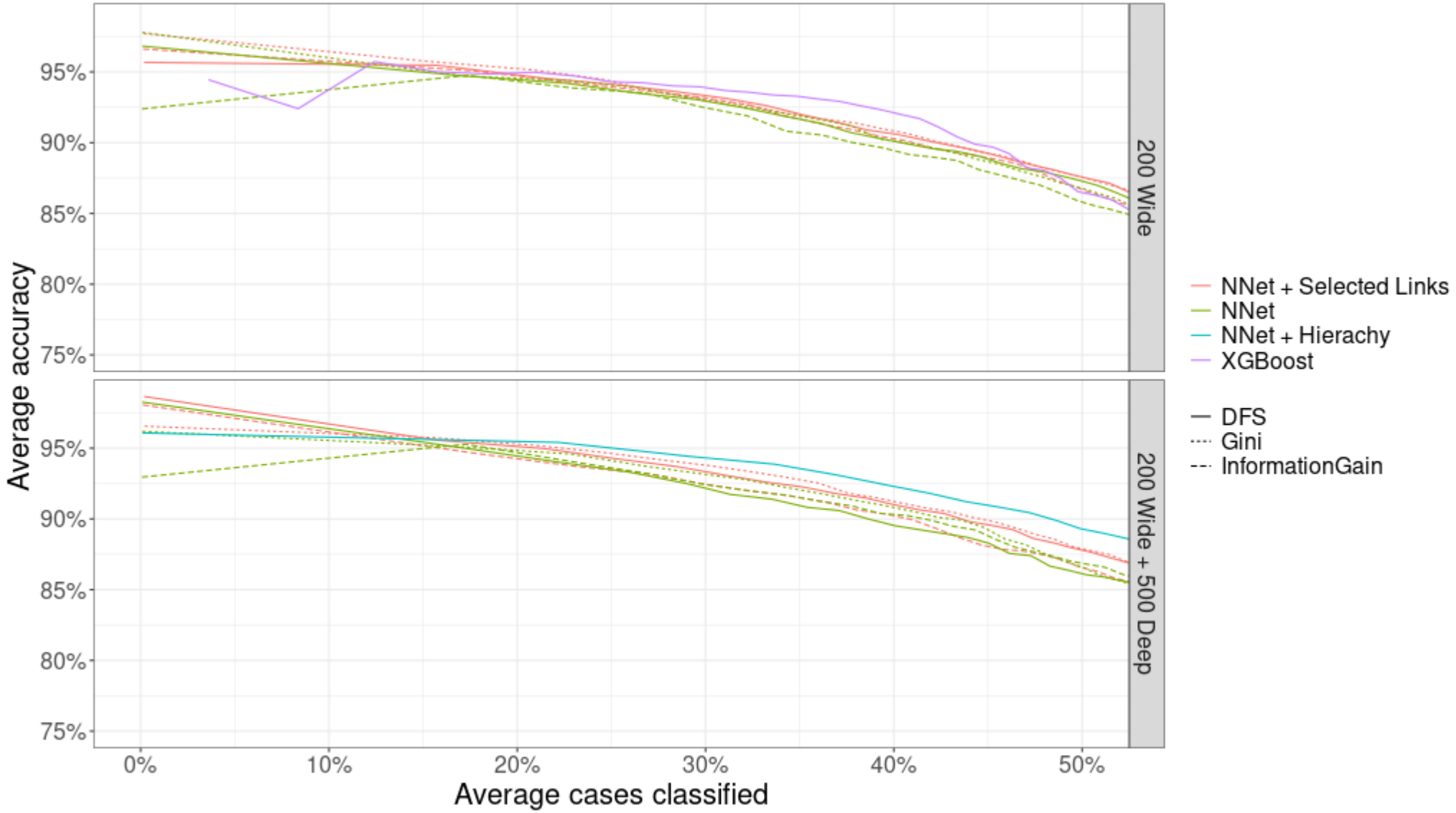


- XGBoost (R-package xgboost)
  - Good out of bag prediction model
  - Not so many tuning parameters

Results - Overall



Results – directly classify



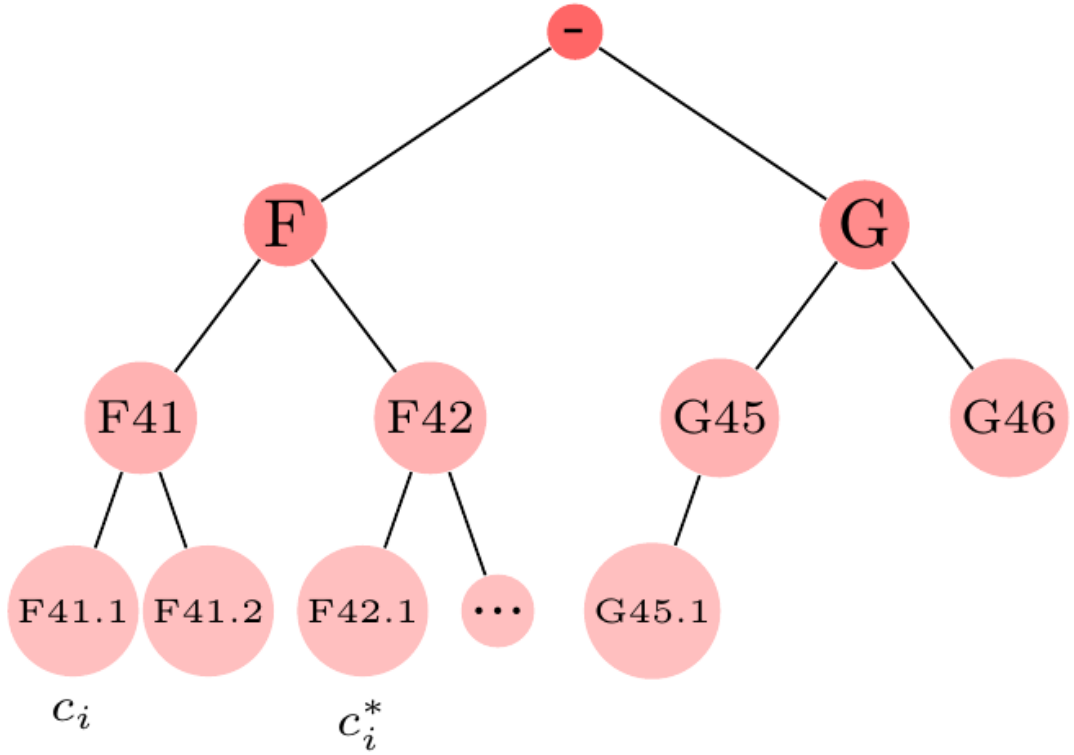
Hierachical Performance measure

- Utilize hierarchical nature of NACE to assess model performance
  - Flat hierarchy assumed up until now

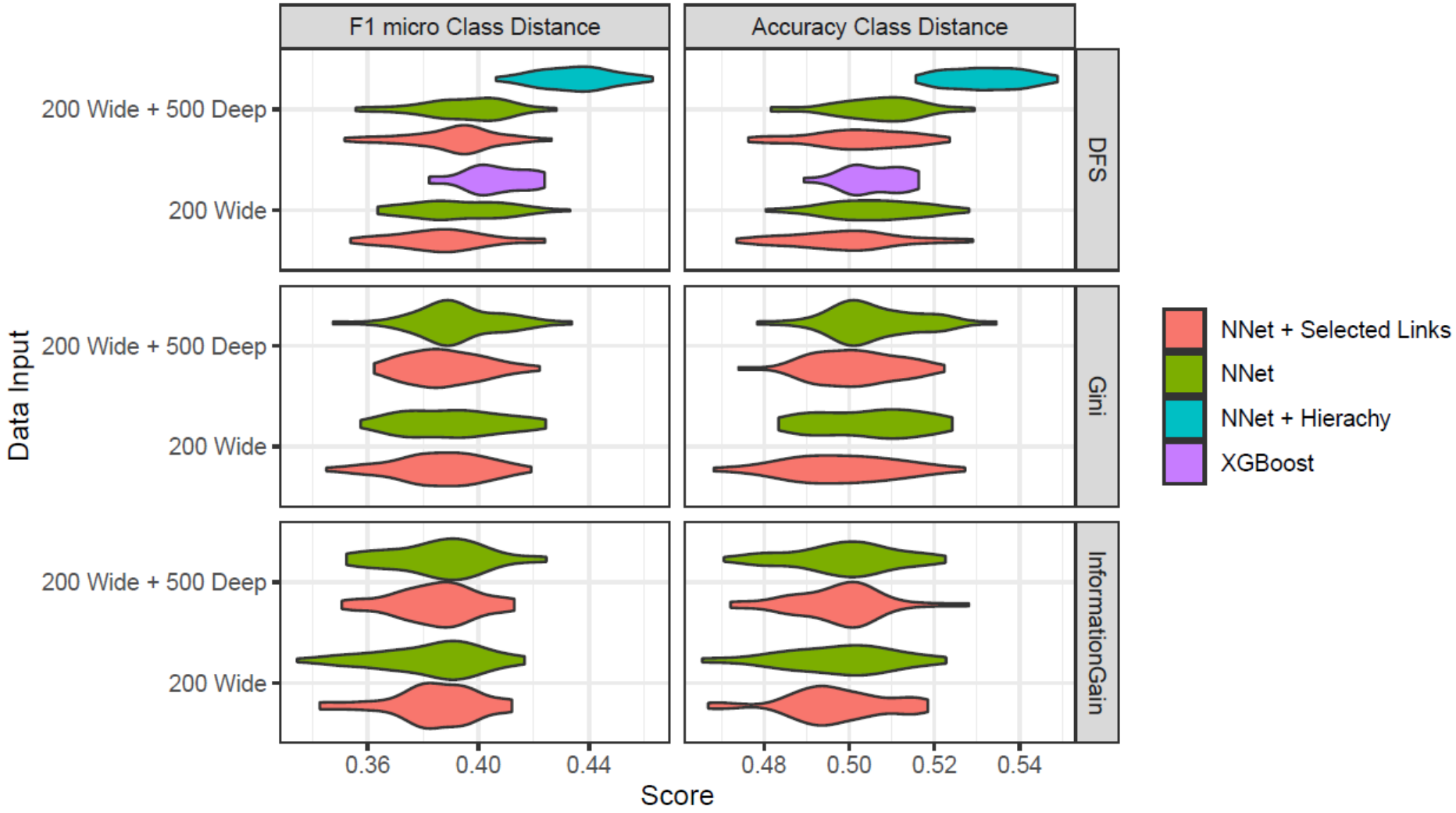
- Class distnace base hierarchical measuer by Sun and Lim (2001)

$$Dis(c_i, c_i^*, L) = 2 \cdot \left(\sum_{l=1}^{L-1} \mathbb{1}_{c^*_{l;i} \neq c_{l;i}} \right)$$

- Define Accuracy, Precision, Recall, F1-Score, …



Results class distance based measures



Result from classifying up to NACE 4 digit

NNet + Hierarchy vs XGBoost iterative predictions

Method	Level	Accuracy	F1	Accuracy (Class Distance)	F1 (Class Distance)	Top k (3/5)
XGBoost	NACE1	0.80	0.80	0.60	0.51	
	NACE2	0.66	0.66	0.46	0.50	
	NACE3	0.55	0.55	0.37	0.36	
	NACE4	0.50	0.50	0.34	0.32	
NNet+Hierarchy	NACE1	0.81	0.81	0.62	0.53	0.95
	NACE2	0.74	0.74	0.57	0.48	0.92
	NACE3	0.66	0.64	0.51	0.42	0.87
	NACE4	0.61	0.59	0.47	0.39	0.82

Conclusions

Conclusions

- NACE classification using text from enterprise websites very challenging
  - Choice of classification method seems to play a smaller role
  - How data is collected and processed seems to be more important
- Direct classification yields unsatisfactory quality
- Supporting manual annotation seems feasible

Rückfragen bitte an

Questions?

STATISTIK AUSTRIA
Guglgasse 13, 1110 Wien