

Free Text Classification with Neural Networks: Training, Process Integration and Results for ISCO-08 Job Titles

Patrick Mertes ▪ Inspirient
Sophie Tschersich ▪ Verian Group

General Online Research Conference 2024 (GOR 24)
Cologne, 22 February 2024



The Need to Automate Free Text Classification

Relevance and Research Question

Why is it necessary to classify job titles?

- **Essential Data:** Job titles are gathered in nearly all empirical social research studies
- **Analytical Necessity:** Classifying occupational groups can be crucial for analyzing survey data
- **Standardization Systems:** Two common systems are used for occupation coding in Germany:
 - **ISCO (international):** Ensures international comparability, governed by the ILO
 - **KldB (national):** Provides national comparability, administered by the Federal Statistical Office

How was this done without neural networks?


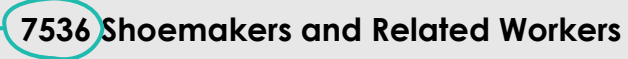
- **Traditional Approach:** Utilized lookup tables for exact matching with official directories
- **Inherent Limitations:** Variations in spelling, paraphrasing, and typos required extensive manual review

**A time-consuming and labour-intensive task
that easily suffers from inconsistencies due to differences in expertise**

ISCO-08 in a Nutshell

Quick Example

ISCO-08: Framework for organizing information on labour and jobs globally, with worldwide acceptance for categorizing job titles

Hierarchical Structure	Count	Example 1	Example 2
		Family medical practitioner	Shoemaker
Major Groups	10	2 - Professionals	7 - Craft and related trades workers
Sub-Major Groups	43	22 - Health Professionals	75 - Food processing, wood working, garment and other craft and related trades workers
Minor Groups	130	221 - Medical Doctors	753 - Garment and Related Trades Workers
Unit Groups	436	2211 - Generalist Medical Practitioners 2212 - Specialist Medical Practitioners	7531 - Tailors, Dressmakers, Furriers and Hatters 7532 - Garment and Related Patternmakers and Cutters 7533 - Sewing, Embroidery and Related Workers 7534 - Upholsterers and Related Workers 7535 - Pelt Dressers, Tanners and Fellmongers
			

Aiming for Efficiency and Consistency

Goals

Goal
Increased efficiency
Increased consistency through recommendations
Retain highest accuracy
Integration into existing workflow
Less than 50% recommended codes need manual decision
More stable project planning and calculation

From Job Titles to Training Data

Machine Learning Setup

Available Data

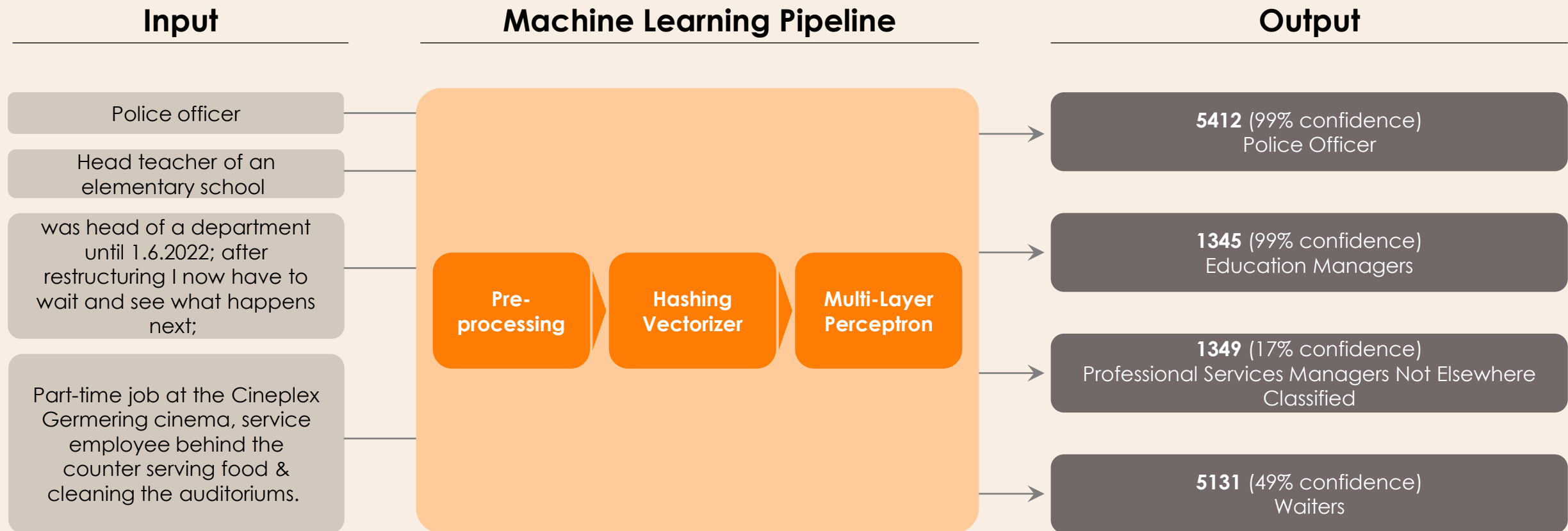
- Historical data from several past years projects
- Around 100.000 distinct German job titles

Difficulties for Machine Learning (examples)

- **Unique:** *Teilzeitarbeitsverhältnis im Kino Cineplex Germering, Service-Mitarbeiter hinter der Theke zum Essen ausgeben & Säüle säubern*
- **Very similar, but different class:** *Lehrerin, Fahrlehrer BE, Grundschullehrerin, Förderlehrerin, Klavierlehrerin, etc.*
- **Multiple jobs with different classes:** *1. Fitnesstrainerin (Aerobickurse) 2. Gymnasiallehrerin (Sport und Deutsch), Schulleitung und Lehrerin Grundschule - Verwaltungsaufgaben, etc.*
- **Gender-specific spelling variations:** *Lehrender, Lehrerin, Lehrer; Kaufmann, Kauffrau; Koch, Köchin*

AI Classification Steps

Technical Dataflow Pipeline



Evaluation of Classification Accuracy

Focus on pre-trained classifiers

Approach

- 1) Train classifiers with selection of Verian's historic manual classifications
- 2) Apply classifier to other datasets to test real-world applicability¹

Classification System

ISCO-08

International Standard
Classification of
Occupations

Verian GER Evaluation Datasets

- Project_A.sav

Accuracy 83%

Coverage 86%

- Project_B.sav

Accuracy 83%

Coverage 84%

- Project_C.sav

Accuracy 83%

Coverage 88%

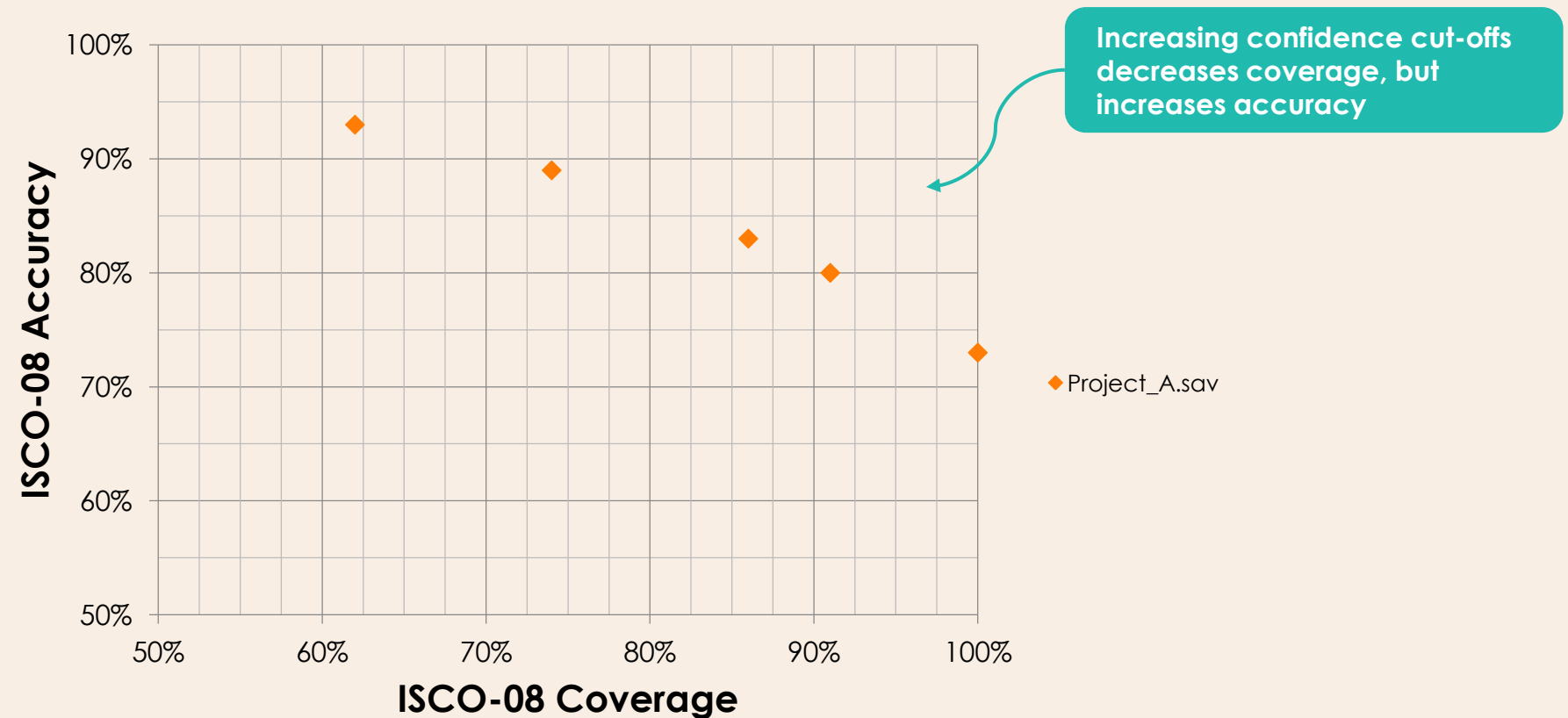
86% of the data
with **83 % accuracy**

But the labels should be 100% accurate. How do we know when to trust the AI?

1. Classifiers configured to only report classifications with >40% confidence for this evaluation

Deep-dive: Accuracy vs. Coverage

Trading lower coverage for increased accuracy



For Real Efficiency Gains, Classifier Results Are Tightly Integrated into Existing Workflow

Inspirit - ISCO-08 Classification Sample.xlsx - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

B13 139424

ISCO-08 Classification
ISCO08_sample.xlsx

89.7%

Recommended Actions

- 1 - Accept Accept without further checking
- 2 - Confirm Confirm that Best Classification is correct
- 3 - Consider Consider having a look at Alternative / Second Alternative Classification
- 4 - Select Select the most appropriate out of all three suggestions
- 5 - Review Review suggestions in detail, fall back to a manual classification if needed
- 6 - Classify Manually perform classification, ignoring suggestions

Input Data		Your Final Classification	AI Recommendation	Best Classification			Alternative Classification				
Case ID	Job Title	ISCO 08 ID	Recommended Action	Comment	ISCO 08 ID	Job Title	Definition	Confidence	ISCO 08 ID	Job Title	Definition
139424	Team leader in area, large company	3341	5 - Review		1330	Information and Commu Information anc		48%	1341	Office Supervisors	Office su
79168	teacher at the community school (full deputat)	2330	2 - Confirm	The first 2 digits of the classification code are lik	2330	Secondary Education Te Secondary educ		77%	2320	Vocational Education Te Vocation	
150272	county clerk	3359	2 - Confirm	The first 2 digits of the classification code are lik	3359	Government Regulatory This unit group		97%	3343	Administrative and Exec Administr	
34273	buyer	3323	1 - Accept	Exact historical match	3323	Buyers	Buyers buy goo	100%	3323	Buyers	Buyers b
94383	nanny	5311	1 - Accept	Exact historical match	5311	Child Care Workers	Child care work	100%	5311	Child Care Workers	Child car
6001	employee industrial clerk	4419	2 - Confirm	The first digit of the classification code is likely t	4110	General Office Clerks	General office c	98%	4419	Clerical Support Worker This unit	
56559	janitor building services	5153	2 - Confirm		5153	Building Caretakers	Building caretak	99%	3112	Civil Engineering Techni Civil eng	
62693	engineer permits	2143	5 - Review		3354	Government Licensing C Government lic		66%	2142	Civil Engineers	Civil eng
50131	scaffolders	7119	1 - Accept	Exact historical match	7119	Building Frame and Rela This unit group		100%	7119	Building Frame and Rela This unit	
73515	automotive designer	2144	5 - Review		2144	Mechanical Engineers	Mechanical eng	43%	3115	Mechanical Engineering Mechni	
53483	global energy and environmental coordinator	1349	4 - Select		2143	Environmental Engineer Environmental		87%	2133	Environmental Protectio Environron	
116	retail trainer	5223	1 - Accept		2320	Vocational Education Te Vocational educ		100%	2359	Teaching Professionals I This unit	
29827	dipl. social pedagogue, deputy head of academy (educational in	2635	5 - Review		2635	Social Work and Course Social work and		59%	1345	Education Managers	Educatio
87895	machine setter/operator	7223	4 - Select		7223	Metal Working Machine Metal working r		79%	8131	Chemical Products Plant Chemica	
32890	dispatcher fire department and rescue control center	3343	2 - Confirm		3258	Ambulance Workers	Ambulance wor	99%	1342	Health Service Manager Health se	
32286	certified psychologist educational counseling family counseling	2634	5 - Review		2635	Social Work and Course Social work and		51%	2634	Psychologists	Psycholo
138007	police system administrator	2522	5 - Review		2522	Systems Administrators	Systems admini	57%	5412	Police Officers	Police of
152452	wellness masseur	5142	1 - Accept	Exact historical match	5142	Beauticians and Related Beauticians and		100%	5142	Beauticians and Related Beauticio	
142925	civil engineering.	2142	1 - Accept	Exact historical match	2142	Civil Engineers	Civil engineers	100%	2142	Civil Engineers	Civil eng
107183	project manager in a company offering saas products for car deal	3341	5 - Review		1221	Sales and Marketing Ma Sales and marke		44%	3322	Commercial Sales Repre Commer	
88544	master bricklayer/administrative assistant->at present	3123	5 - Review		3343	Administrative and Exec Administrative		45%	3359	Government Regulatory This unit	
117560	transport clerk	4323	3 - Consider		4323	Transport Clerks	Transport clerks	92%	9333	Freight Handlers	Freight h
120346	locksmith in underground mining	7212	3 - Consider		8111	Miners and Quarriers	Miners and quar	94%	7222	Toolmakers and Relatec Toolmak	
39275	educational consultant and research associate	3412	4 - Select		2310	University and Higher Ec University and H		87%	2351	Education Methods Spe Educatio	
130418	special needs teacher in inclusion at a sek. 1 (main school)	2352	3 - Consider		2352	Special Needs Teachers	Special needs te	92%	3412	Social Work Associate Pi Social wo	
51707	managing director guidance scientific advisory board of the conf	1111	5 - Review	The first 2 digits of the classification code are lik	1120	Managing Directors and Managing direct		60%	1114	Senior Officials of Spec Senior of	

incl. per-interview recommended action and suggested classifications

B13 139424

ISCO-08 Classification

ISCO08_sample.xlsx

89.7%

Recommended Actions

- 1 - Accept Accept without further checking
- 2 - Confirm Confirm that Best Classification is correct
- 3 - Consider Consider having a look at Alternative / Second Alternative Classification
- 4 - Select Select the most appropriate out of all three suggestions
- 5 - Review Review suggestions in detail, fall back to a manual classification if needed
- 6 - Classify Manually perform classification, ignoring suggestions

Input Data		Your Final Classification	AI Recommendation		Best Classification			
Case ID	Job Title	ISCO 08 ID	Recommended Action	Comment	ISCO 08 ID	Job Title	Definition	Confidence
139424	team leader it area, large company	3341	5 - Review		1330	Information and Comm	Information and	48%
79168	teacher at the community school (full deputat)	2330	2 - Confirm	The first 2 digits of the classification code are lik	2330	Secondary Education Te	Secondary educ	97%
150272	county clerk	3359	2 - Confirm	The first 2 digits of the classification code are lik	3359	Government Regulatory	This unit group	97%
34273	buyer	3323	1 - Accept	Exact historical match	3323	Buyers	Buyers buy goo	100%
94383	nanny	5311	1 - Accept	Exact historical match	5311	Child Care Workers	Child care work	100%
6001	employee industrial clerk	4419	2 - Confirm	The first digit of the classification code is likely t	4110	General Office Clerks	General office c	98%
56559	janitor building services	5153	2 - Confirm		5153	Building Caretakers	Building caretak	99%
62693	engineer permits	2141	5 - Review		3354	Government Licensing C	Government lic	66%
50131	scaffolders	7119	1 - Accept	Exact historical match	7119	Building Frame and Rel	This unit group	100%
73515	automotive designer	2144	5 - Review		2144	Mechanical Engineers	Mechanical eng	43%
53483	global energy and enviromental coordinator	1349	4 - Select		2143	Environmental Engineer	Environmental e	87%
116	* retail trainer	5223	1 - Accept		2320	Vocational Education Te	Vocational educ	100%
29827	dipl. social pedagogue, deputy head of academy (educational in	2635	5 - Review		2635	Social Work and Course	Social work and	59%
87895	machine setter/operator	7223	4 - Select		7223	Metal Working Machine	Metal working r	79%
32890	dispatcher fire department and rescue control center	3343	2 - Confirm		3258	Ambulance Workers	Ambulance wor	99%
32286	certified psychologist educational counseling family counseling	2634	5 - Review		2635	Social Work and Course	Social work and	51%
138007	police system administrator	2522	5 - Review		2522	Systems Aministrators	Systems admini	57%
152452	wellness masseur	5142	1 - Accept	Exact historical match	5142	Beauticians and Related	Beauticians and	100%
142925	civil engineering.	2142	1 - Accept	Exact historical match	2142	Civil Engineers	Civil engineers	100%
107183	project manager in a company offering saas products for car deal	3341	5 - Review		1221	Sales and Marketing Ma	Sales and marke	44%
88544	master bricklayer/administrative assistant->at present	3123	5 - Review		3343	Administrative and Exec	Administrative	45%

Six Key Benefits

Business Impact

Goal	
Increased efficiency	✓
Increased consistency through recommendations	✓
Retain highest accuracy	✓
Integration into existing workflow	✓
Less than 50% recommended codes need manual decision	✓
More stable project planning and calculation	✓

Manual decisions still required, also as per client mandate, but much faster

Future Work

Next steps to further advance this method

Potential improvements to NN classifiers:

- **Continuous Improvement:** Accuracy enhancement through ongoing learning from new data and corrections.
- **Training for Coders:** Up-skilling manual coders with datasets for improved efficiency and accuracy.
- **Broader Implementation:** Extending the model to additional national and international classifications, like ISCED and KIdB2010.

First tests show that the same approach works very well.

KIdB 2010

Klassifikation der Berufe

Accuracy 82%

Coverage 89%

WZ 2008

Klassifikation der Wirtschaftszweige

Accuracy 89%

Coverage 90%

Contact Information

Patrick Mertes ▪ patrick.mertes@inspirient.com

Sophie Tschersich ▪ sophie.tschersich@veriangroup.com

