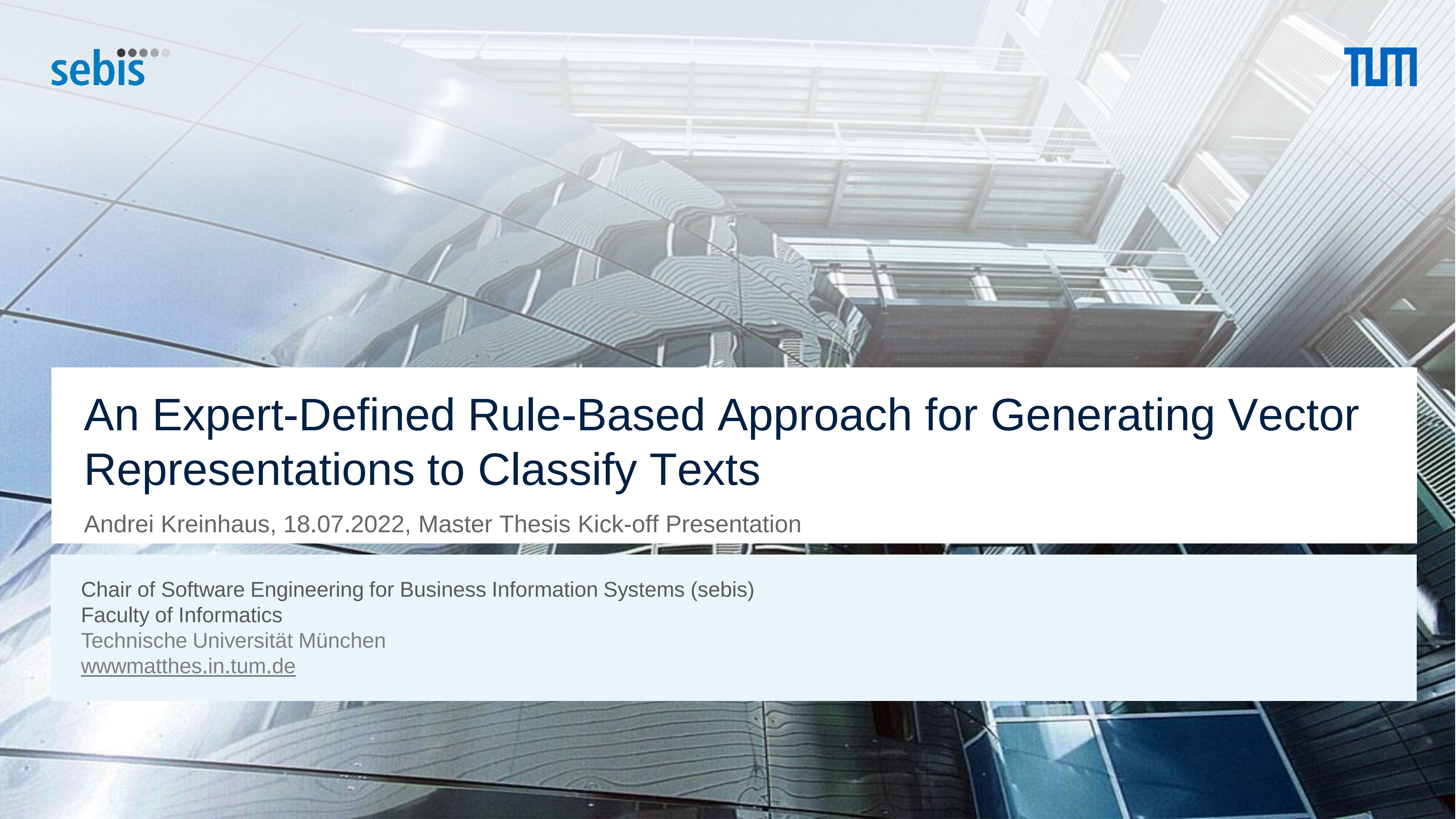


An Expert-Defined Rule-Based Approach for Generating Vector Representations to Classify Texts

Andrei Kreinhaus, 18.07.2022, Master Thesis Kick-off Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

Introduction

- Motivation
- Challenges
- Approach
- Objectives

Research concept

- Research questions
- Methodology

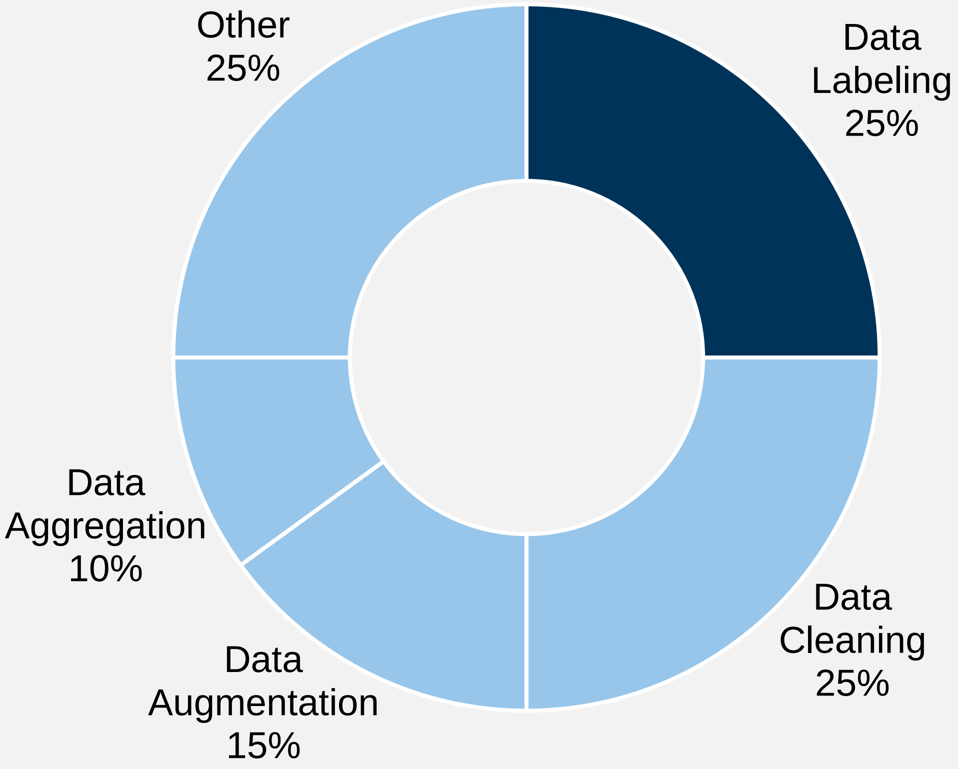
Dataset

Progress

Timeline

Motivation

A quarter of time is spent on manual labeling in ML related activities



For 25 classes **over 40 working hours** are required to create training data

Assumptions

- At least 100 labeled documents per class
- Average reading speed is 200 words/min
- Average complaint length is around 100 words
- Considering 30 seconds for multiclass labeling, 1 minute per complaint in average is required

Unresolved issues

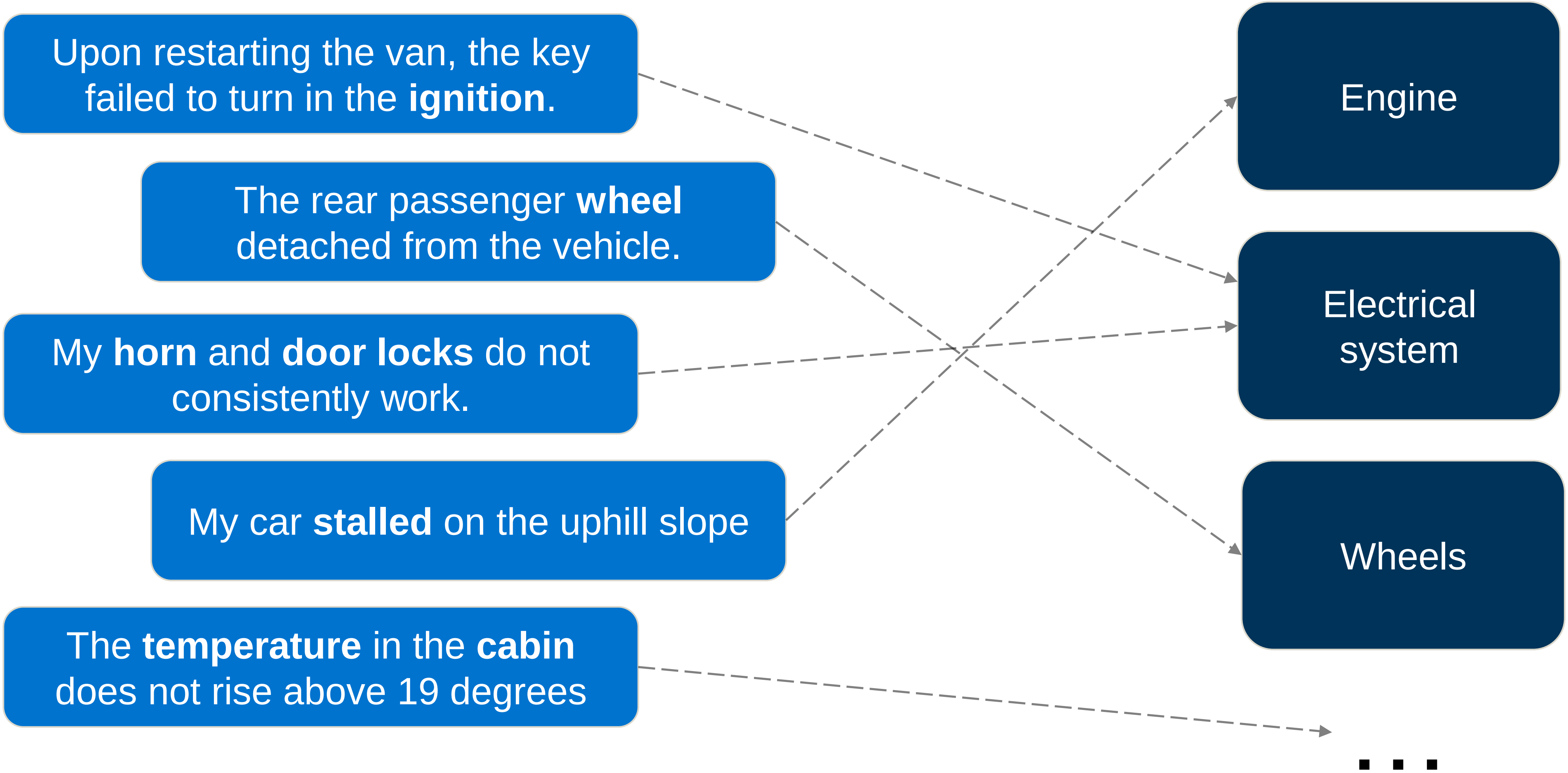
- The approach is prone to human errors
- Human concentration is not constantly high
- Limitations of the supervised learning

https://www.cognilytica.com/document/data-preparation-labeling-for-ai-2020/

Challenges I – Classification of Automotive Complains by Car Component

Customer complaints

25 classes = 25 car components



Upon restarting the van, the key failed to turn in the **ignition**.

The rear passenger **wheel** detached from the vehicle.

My **horn** and **door locks** do not consistently work.

My car **stalled** on the uphill slope

The **temperature** in the **cabin** does not rise above 19 degrees

Engine

Electrical system

Wheels

...

Challenges II – Review of Classification Challenges in Supervised Learning

Data Perspective

- Zero-shot/Few-shot learning
- Special domain with many terminologies
- The multi-label text classification task
- Labeling is time-consuming
- Low flexibility in case of an objective change

Model Perspective

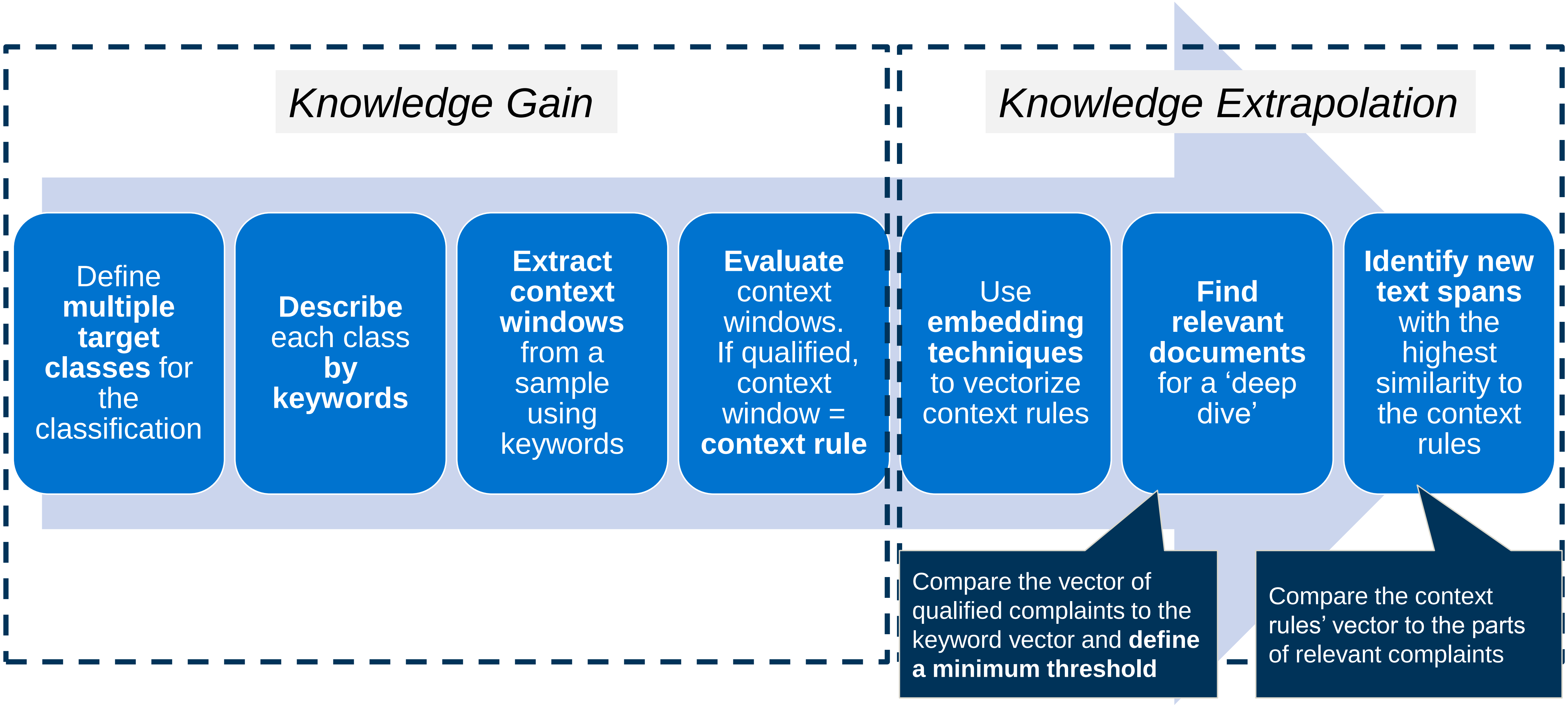
- Text representation
- Model integration
- Model efficiency

Performance Perspective

- The semantic robustness of the model
- The interpretability of the model

Source: Kammoun et al. 2022

Approach – Combined approach of multiple NLP techniques



Knowledge Gain

Define **multiple target classes** for the classification

Describe each class **by keywords**

Extract context windows from a sample using keywords

Evaluate context windows. If qualified, context window = **context rule**

Knowledge Extrapolation

Use **embedding techniques** to vectorize context rules

Find relevant documents for a ‘deep dive’

Identify new text spans with the highest similarity to the context rules

Compare the vector of qualified complaints to the keyword vector and **define a minimum threshold**

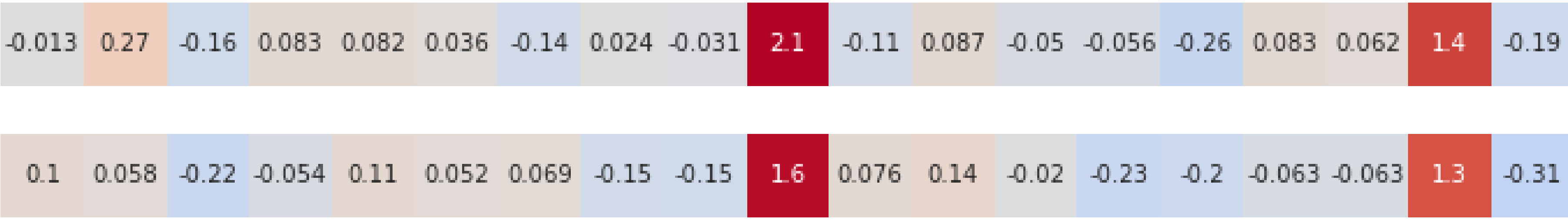
Compare the context rules’ vector to the parts of relevant complaints

Approach – Example

Classes = {…, ‘exterior light’: [‘light’, ‘head light’, ‘fog light’, ‘**high beam**’, …, ‘reversing light’], …}

first_sentence = "My **high beam** suddenly comes on on the highway“

second_sentence = "My *headlights* started flashing as I accelerated“



-0.013	0.27	-0.16	0.083	0.082	0.036	-0.14	0.024	-0.031	2.1	-0.11	0.087	-0.05	-0.056	-0.26	0.083	0.062	1.4	-0.19	encode(first_sentence)
0.1	0.058	-0.22	-0.054	0.11	0.052	0.069	-0.15	-0.15	1.6	0.076	0.14	-0.02	-0.23	-0.2	-0.063	-0.063	1.3	-0.31	encode(second_sentence)

$$cosine\ similarity = \frac{x * y}{\|x\| * \|y\|}$$

$$cosine\_similarity(encode(first\_sentence), encode(second\_sentence)) = 0.782$$

Consider the context window if the cosine similarity larger than a chosen threshold.

Objectives

Observing and documenting **challenges throughout the process**

E.g., data quality, vector quality, training time, bottlenecks, etc.

Comparing different vectorization techniques from TF-IDF to BERT

Suggesting an optimal combination of different NLP techniques

Comparing results **to unsupervised NLP techniques**

E.g., clustering or topic modeling

Creating a **framework** for multiclass classification based on **predefined classes with keywords** to **reduce labeling time** and produce a **structured dataset** as a result

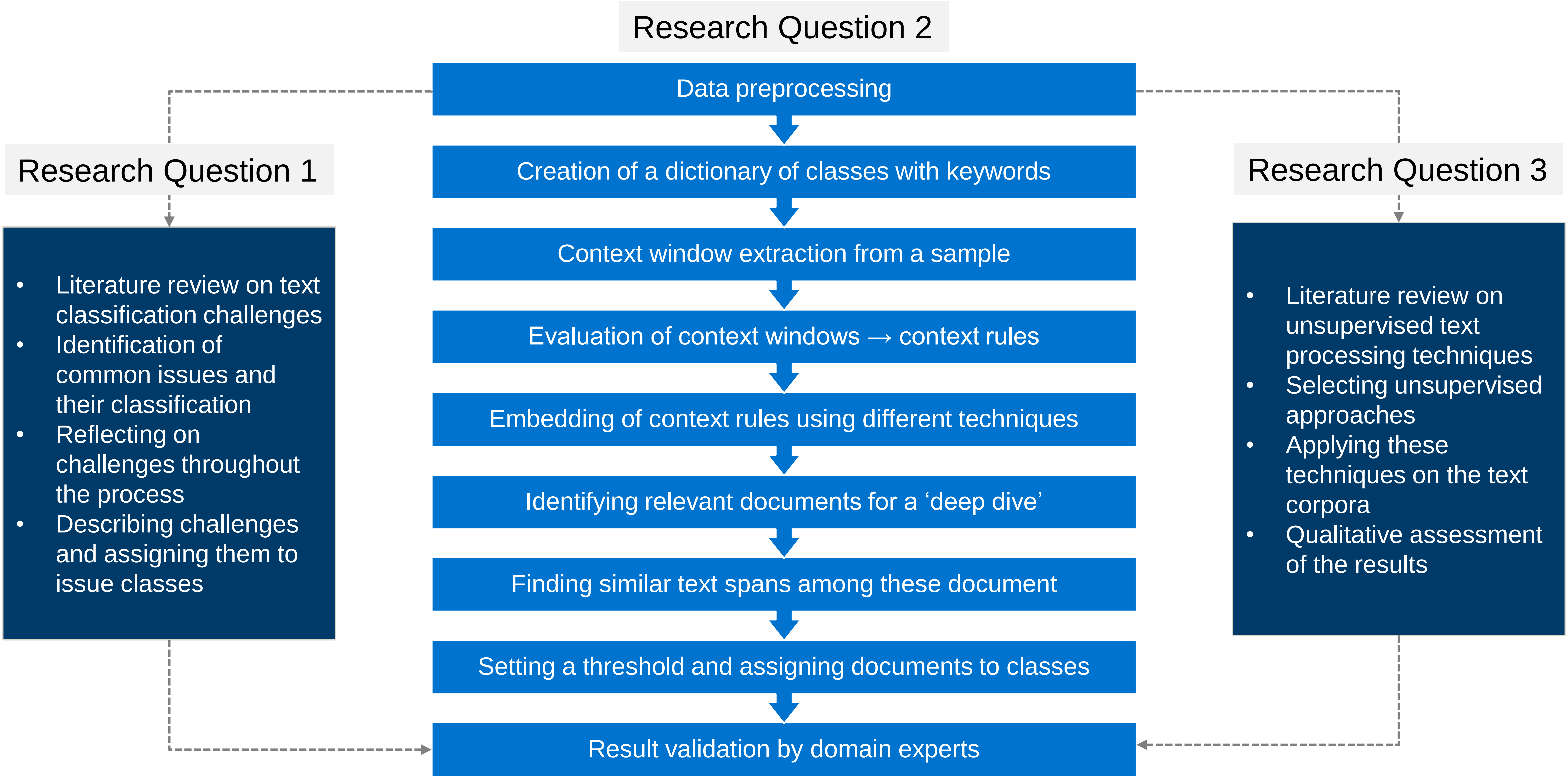
Research Questions

1. **What are the challenges** faced when trying to create structured datasets from unstructured documents?

2. **Which NLP methods** can be **combined** with domain expertise to facilitate the extrapolation from context rules to training data?

3. How do these novel methods **compare to** current methods of **unsupervised learning** in the context of automotive customer data?

Methodology



Research Question 2

Data preprocessing

Creation of a dictionary of classes with keywords

Context window extraction from a sample

Evaluation of context windows → context rules

Embedding of context rules using different techniques

Identifying relevant documents for a ‘deep dive’

Finding similar text spans among these document

Setting a threshold and assigning documents to classes

Result validation by domain experts

Research Question 1

- Literature review on text classification challenges
- Identification of common issues and their classification
- Reflecting on challenges throughout the process
- Describing challenges and assigning them to issue classes

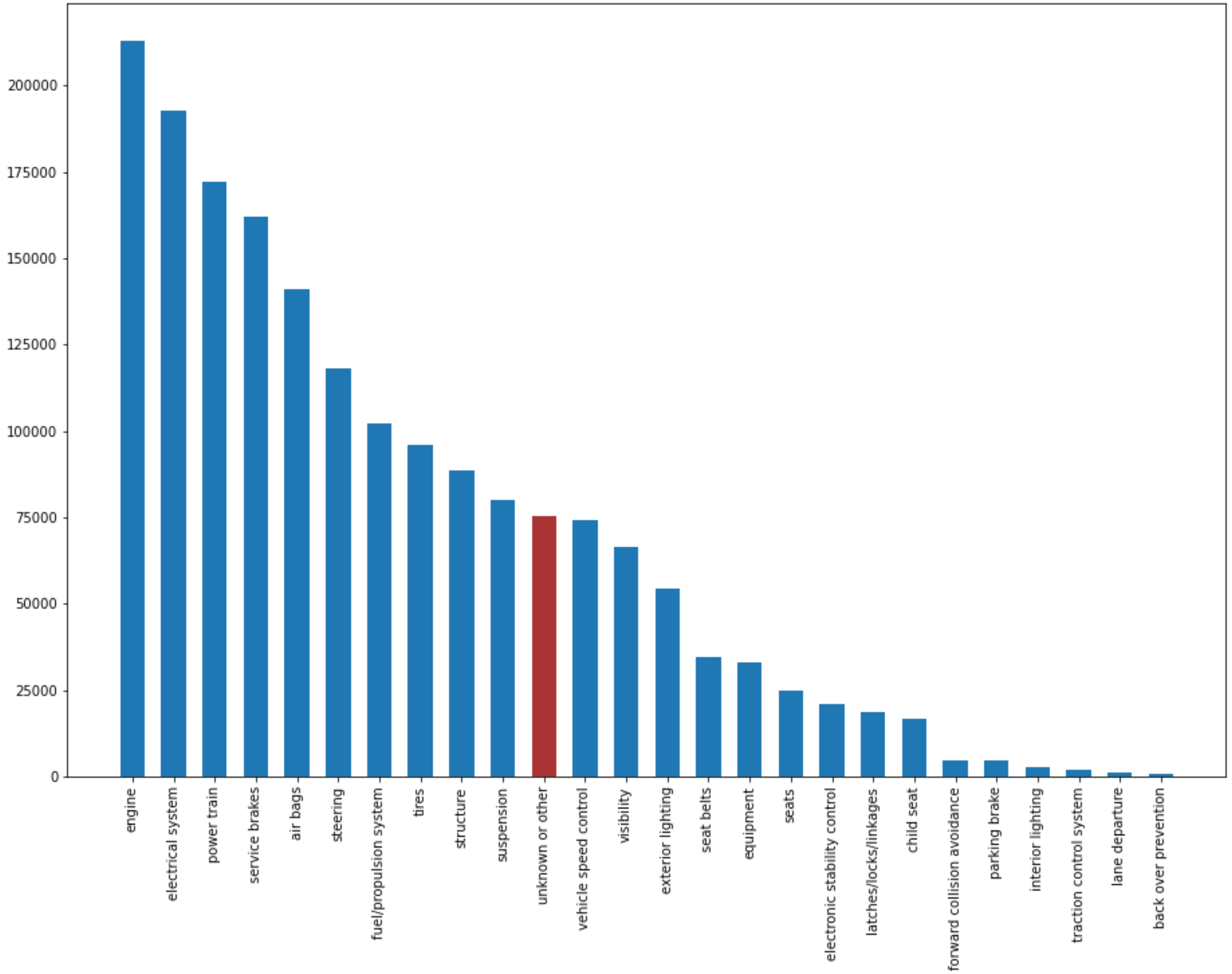
Research Question 3

- Literature review on unsupervised text processing techniques
- Selecting unsupervised approaches
- Applying these techniques on the text corpora
- Qualitative assessment of the results

Dataset – Description and Number of Complaints per Class

- **Anonymized open-source dataset** of the U.S. Department of Transportation (NHTSA)
- **Identification of safety issues** and determining if a safety-related defect trend exists
- Subjects considered: **Vehicles, Tires, Child Safety Seats, Equipment**
- **Over 1.3 Mio unique complaints** updated on daily basis
- Some **complaints have multiple labels**
- **Highly imbalanced dataset**
- **Over 75000 unclassified complaints**

Number of complaints per class

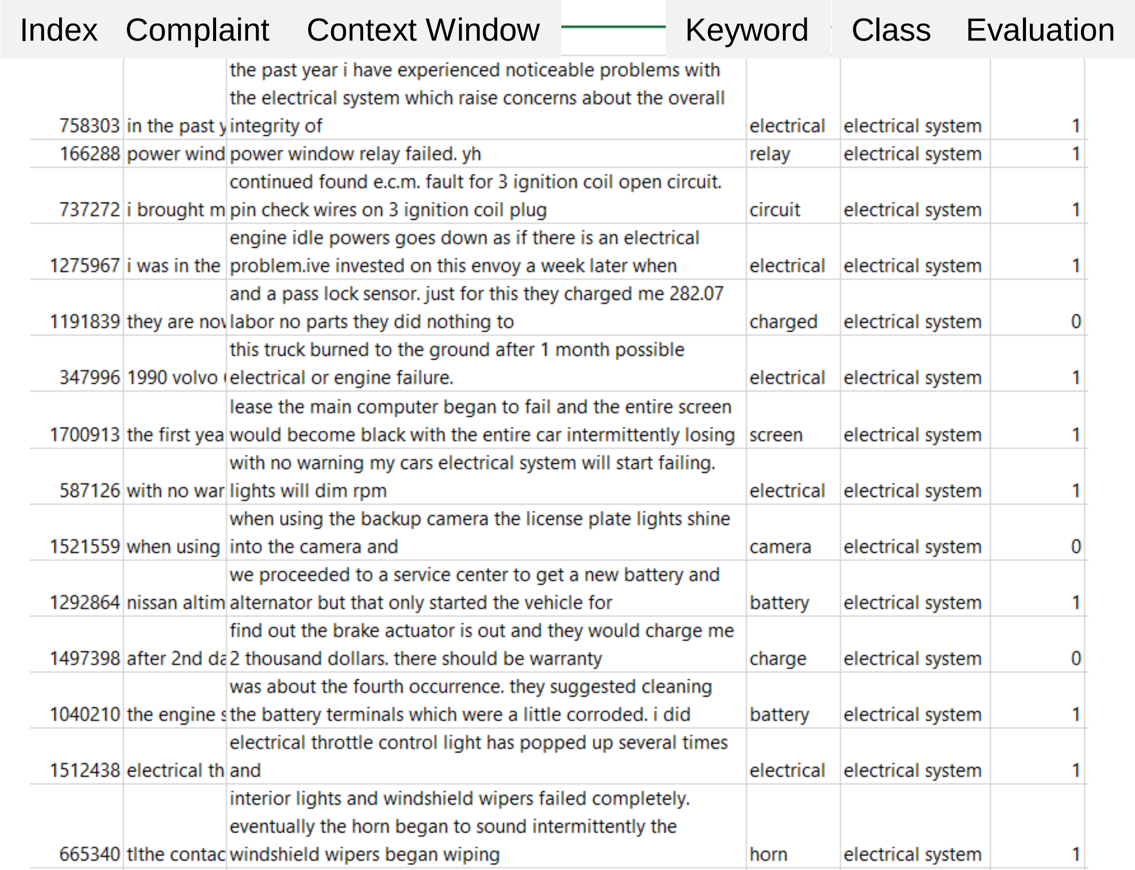


Progress – Challenges of the Chosen Approach

	Processing Steps	Challenges	Comments
1	Data preprocessing	• Preserve readability	Manual evaluation
2	Define multiple target classes for the classification	• Number of classes • Hierarchical dependency	Usually defined in a task
3	Describe each class by keywords	• Number of keywords • Precision vs simplicity trade-off	In coordination with domain experts
4	Extract context windows from a sample using keywords	• Length of the context window	As tokens or sentences
5	Evaluate context windows. If qualified, a context window becomes a context rule	• Manual evaluation • Ambiguous context windows	A few words may be missing
6	Use embedding techniques to vectorize context rules	• Selecting embedding technique • Complexity vs vector quality trade-off	Experimenting is time consuming
7	Find relevant documents for a ‘deep dive’. Compare the vector of qualified complaints to the keyword vector and define a minimum threshold	• Using initial keywords vs keywords used for context windows only • A general threshold vs one for each class	Experimenting is time consuming
8	Identify new text spans with the highest similarity to the context rules. Compare the context rules’ vector to the parts of relevant complaints	• Keep complexity within limits • Sliding window vs tokens vs sentences • Choosing a threshold	Experimenting is time consuming

Progress – Example of Evaluated Context Windows

Evaluation of identified context windows

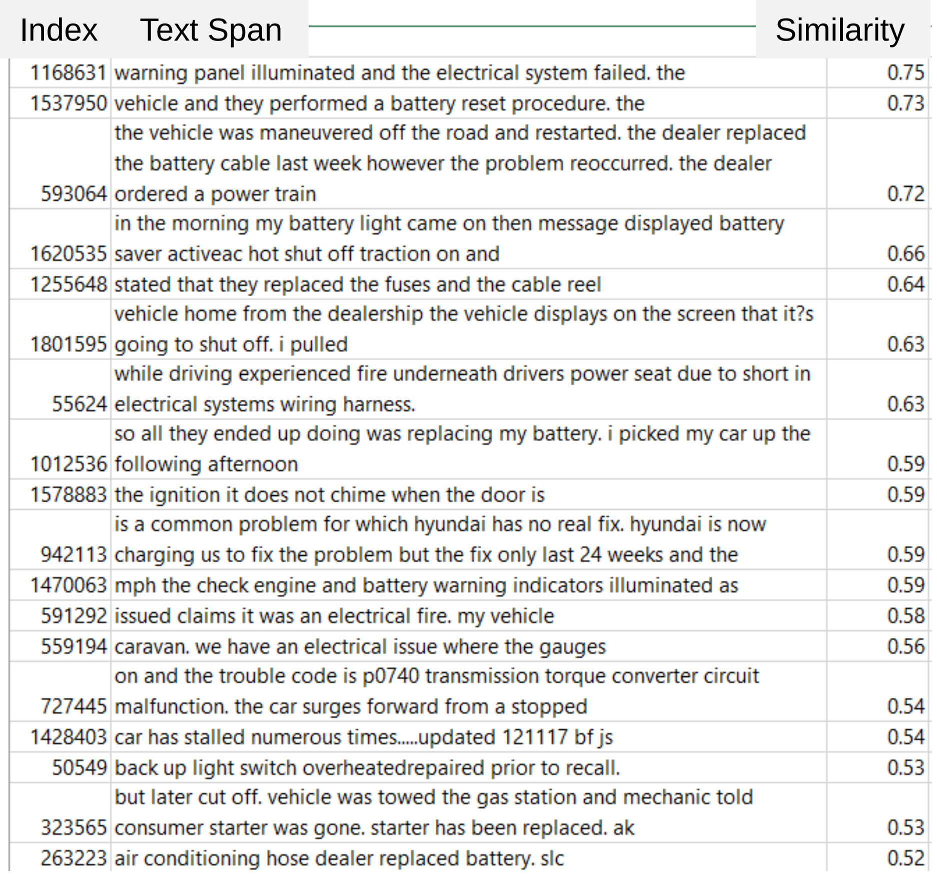


Index	Complaint	Context Window	Keyword	Class	Evaluation
758303	in the past y	the past year i have experienced noticeable problems with the electrical system which raise concerns about the overall integrity of	electrical	electrical system	1
166288	power wind	power window relay failed. yh	relay	electrical system	1
737272	i brought m	continued found e.c.m. fault for 3 ignition coil open circuit. pin check wires on 3 ignition coil plug	circuit	electrical system	1
1275967	i was in the	engine idle powers goes down as if there is an electrical problem.ive invested on this envoy a week later when	electrical	electrical system	1
1191839	they are nov	and a pass lock sensor. just for this they charged me 282.07 labor no parts they did nothing to	charged	electrical system	0
347996	1990 volvo	this truck burned to the ground after 1 month possible electrical or engine failure.	electrical	electrical system	1
1700913	the first yea	lease the main computer began to fail and the entire screen would become black with the entire car intermittently losing	screen	electrical system	1
587126	with no war	with no warning my cars electrical system will start failing. lights will dim rpm	electrical	electrical system	1
1521559	when using	when using the backup camera the license plate lights shine into the camera and	camera	electrical system	0
1292864	nissan altim	we proceeded to a service center to get a new battery and alternator but that only started the vehicle for	battery	electrical system	1
1497398	after 2nd da	find out the brake actuator is out and they would charge me 2 thousand dollars. there should be warranty	charge	electrical system	0
1040210	the engine s	was about the fourth occurrence. they suggested cleaning the battery terminals which were a little corroded. i did	battery	electrical system	1
1512438	electrical th	electrical throttle control light has popped up several times and	electrical	electrical system	1
665340	tlthe contac	interior lights and windshield wipers failed completely. eventually the horn began to sound intermittently the windshield wipers began wiping	horn	electrical system	1

Embedding Techniques

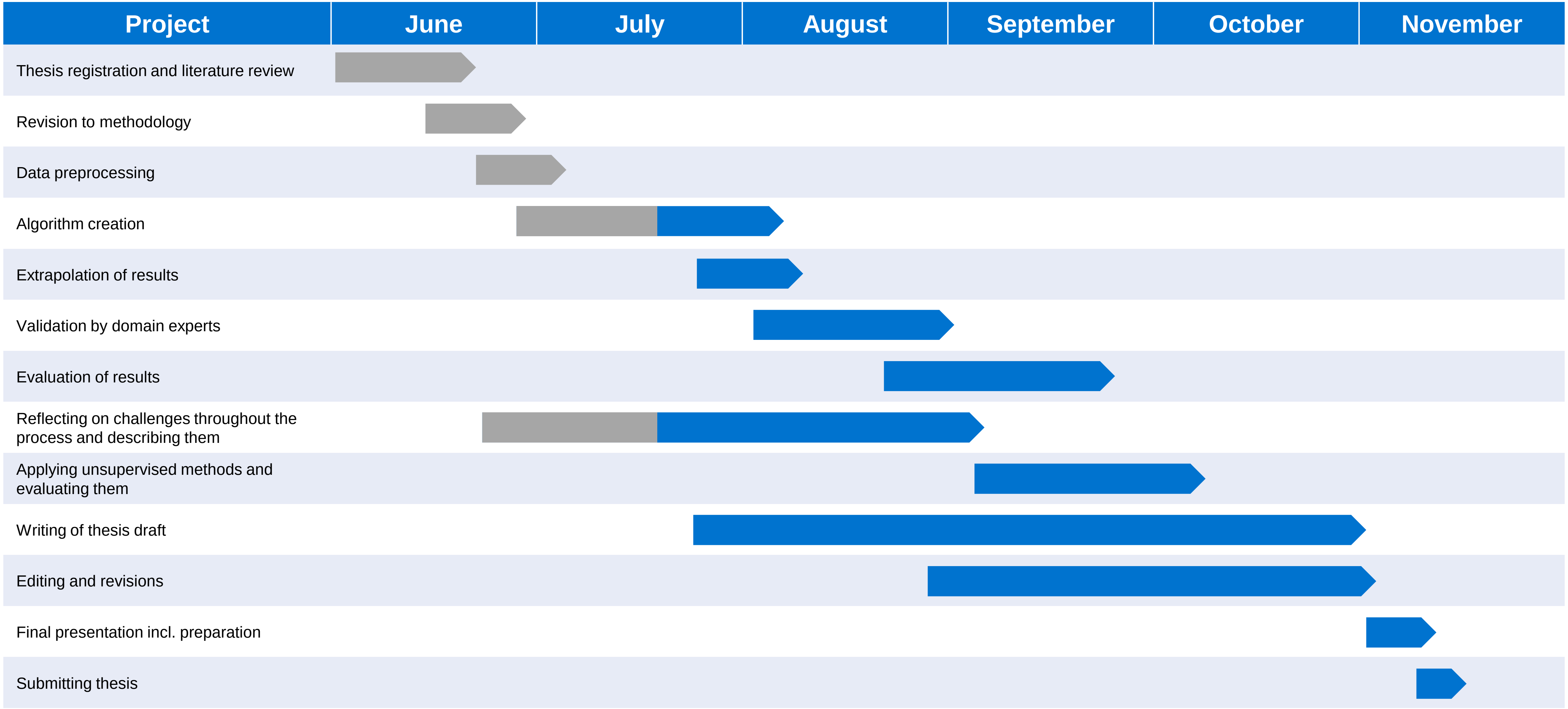
Cosine Similarity

Identification of similar text spans



Index	Text Span	Similarity
1168631	warning panel illuminated and the electrical system failed. the	0.75
1537950	vehicle and they performed a battery reset procedure. the	0.73
593064	the vehicle was maneuvered off the road and restarted. the dealer replaced the battery cable last week however the problem reoccurred. the dealer ordered a power train	0.72
1620535	in the morning my battery light came on then message displayed battery saver activeac hot shut off traction on and	0.66
1255648	stated that they replaced the fuses and the cable reel	0.64
1801595	vehicle home from the dealership the vehicle displays on the screen that it?s going to shut off. i pulled	0.63
55624	while driving experienced fire underneath drivers power seat due to short in electrical systems wiring harness.	0.63
1012536	so all they ended up doing was replacing my battery. i picked my car up the following afternoon	0.59
1578883	the ignition it does not chime when the door is	0.59
942113	is a common problem for which hyundai has no real fix. hyundai is now charging us to fix the problem but the fix only last 24 weeks and the	0.59
1470063	mph the check engine and battery warning indicators illuminated as	0.59
591292	issued claims it was an electrical fire. my vehicle	0.58
559194	caravan. we have an electrical issue where the gauges	0.56
727445	on and the trouble code is p0740 transmission torque converter circuit malfunction. the car surges forward from a stopped	0.54
1428403	car has stalled numerous times.....updated 121117 bf js	0.54
50549	back up light switch overheatedrepaired prior to recall.	0.53
323565	but later cut off. vehicle was towed the gas station and mechanic told consumer starter was gone. starter has been replaced. ak	0.53
263223	air conditioning hose dealer replaced battery. slc	0.52

Timeline



Project | June | July | August | September | October | November

Thesis registration and literature review

Revision to methodology

Data preprocessing

Algorithm creation

Extrapolation of results

Validation by domain experts

Evaluation of results

Reflecting on challenges throughout the process and describing them

Applying unsupervised methods and evaluating them

Writing of thesis draft

Editing and revisions

Final presentation incl. preparation

Submitting thesis

Andrei Kreinhaus
M.Sc. Management & Technology

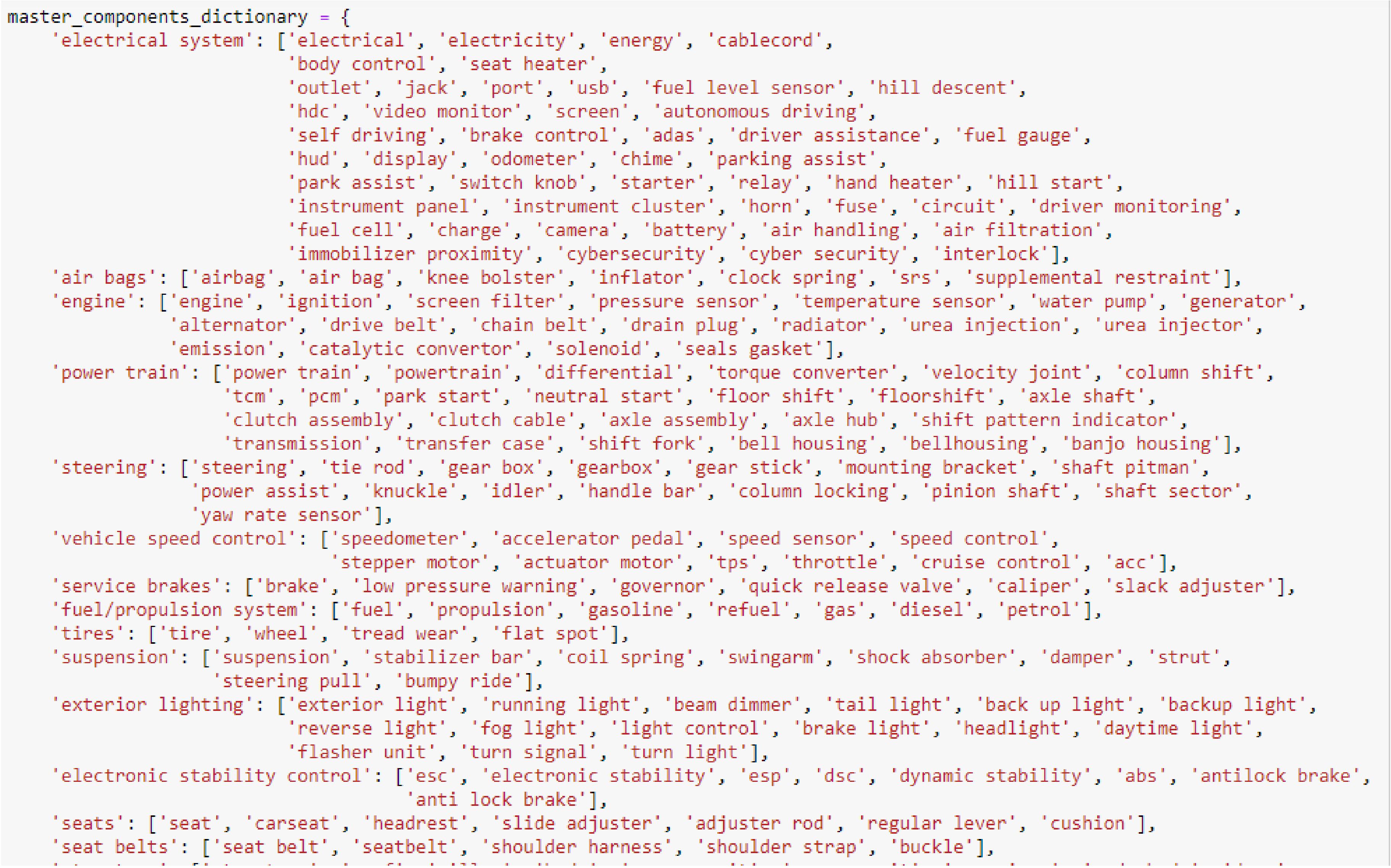
Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17132
Fax +49.89.289.17136

andrei.kreinhaus@tum.de
wwwmatthes.in.tum.de

Back-up: Dictionary of Classes with Keywords



```
master_components_dictionary = {
    'electrical system': ['electrical', 'electricity', 'energy', 'cablecord',
                          'body control', 'seat heater',
                          'outlet', 'jack', 'port', 'usb', 'fuel level sensor', 'hill descent',
                          'hdc', 'video monitor', 'screen', 'autonomous driving',
                          'self driving', 'brake control', 'adas', 'driver assistance', 'fuel gauge',
                          'hud', 'display', 'odometer', 'chime', 'parking assist',
                          'park assist', 'switch knob', 'starter', 'relay', 'hand heater', 'hill start',
                          'instrument panel', 'instrument cluster', 'horn', 'fuse', 'circuit', 'driver monitoring',
                          'fuel cell', 'charge', 'camera', 'battery', 'air handling', 'air filtration',
                          'immobilizer proximity', 'cybersecurity', 'cyber security', 'interlock'],
    'air bags': ['airbag', 'air bag', 'knee bolster', 'inflator', 'clock spring', 'srs', 'supplemental restraint'],
    'engine': ['engine', 'ignition', 'screen filter', 'pressure sensor', 'temperature sensor', 'water pump', 'generator',
               'alternator', 'drive belt', 'chain belt', 'drain plug', 'radiator', 'urea injection', 'urea injector',
               'emission', 'catalytic convertor', 'solenoid', 'seals gasket'],
    'power train': ['power train', 'powertrain', 'differential', 'torque converter', 'velocity joint', 'column shift',
                    'tcm', 'pcm', 'park start', 'neutral start', 'floor shift', 'floorshift', 'axle shaft',
                    'clutch assembly', 'clutch cable', 'axle assembly', 'axle hub', 'shift pattern indicator',
                    'transmission', 'transfer case', 'shift fork', 'bell housing', 'bellhousing', 'banjo housing'],
    'steering': ['steering', 'tie rod', 'gear box', 'gearbox', 'gear stick', 'mounting bracket', 'shaft pitman',
                 'power assist', 'knuckle', 'idler', 'handle bar', 'column locking', 'pinion shaft', 'shaft sector',
                 'yaw rate sensor'],
    'vehicle speed control': ['speedometer', 'accelerator pedal', 'speed sensor', 'speed control',
                              'stepper motor', 'actuator motor', 'tps', 'throttle', 'cruise control', 'acc'],
    'service brakes': ['brake', 'low pressure warning', 'governor', 'quick release valve', 'caliper', 'slack adjuster'],
    'fuel/propulsion system': ['fuel', 'propulsion', 'gasoline', 'refuel', 'gas', 'diesel', 'petrol'],
    'tires': ['tire', 'wheel', 'tread wear', 'flat spot'],
    'suspension': ['suspension', 'stabilizer bar', 'coil spring', 'swingarm', 'shock absorber', 'damper', 'strut',
                   'steering pull', 'bumpy ride'],
    'exterior lighting': ['exterior light', 'running light', 'beam dimmer', 'tail light', 'back up light', 'backup light',
                          'reverse light', 'fog light', 'light control', 'brake light', 'headlight', 'daytime light',
                          'flasher unit', 'turn signal', 'turn light'],
    'electronic stability control': ['esc', 'electronic stability', 'esp', 'dsc', 'dynamic stability', 'abs', 'antilock brake',
                                     'anti lock brake'],
    'seats': ['seat', 'carseat', 'headrest', 'slide adjuster', 'adjuster rod', 'regular lever', 'cushion'],
    'seat belts': ['seat belt', 'seatbelt', 'shoulder harness', 'shoulder strap', 'buckle'],
```