KU LEUVEN

Detecting industry trends using web scraping on Belgian company websites

A quantitative study on CSR reporting of Belgian company websites

Michiel Devos (R0764064)

Nicolas Engelen (R0757765)

Promoter: Prof. Dr. Bart Baesens
Daily Supervisor: Manon Reusens
Statistiek Vlaanderen: Michael Reusens

Problem setting

Problem setting

How to measure the CSR commitment of companies?

- Surveys
- Practical test
- Displayed commitment ← Focus of this research

Research questions

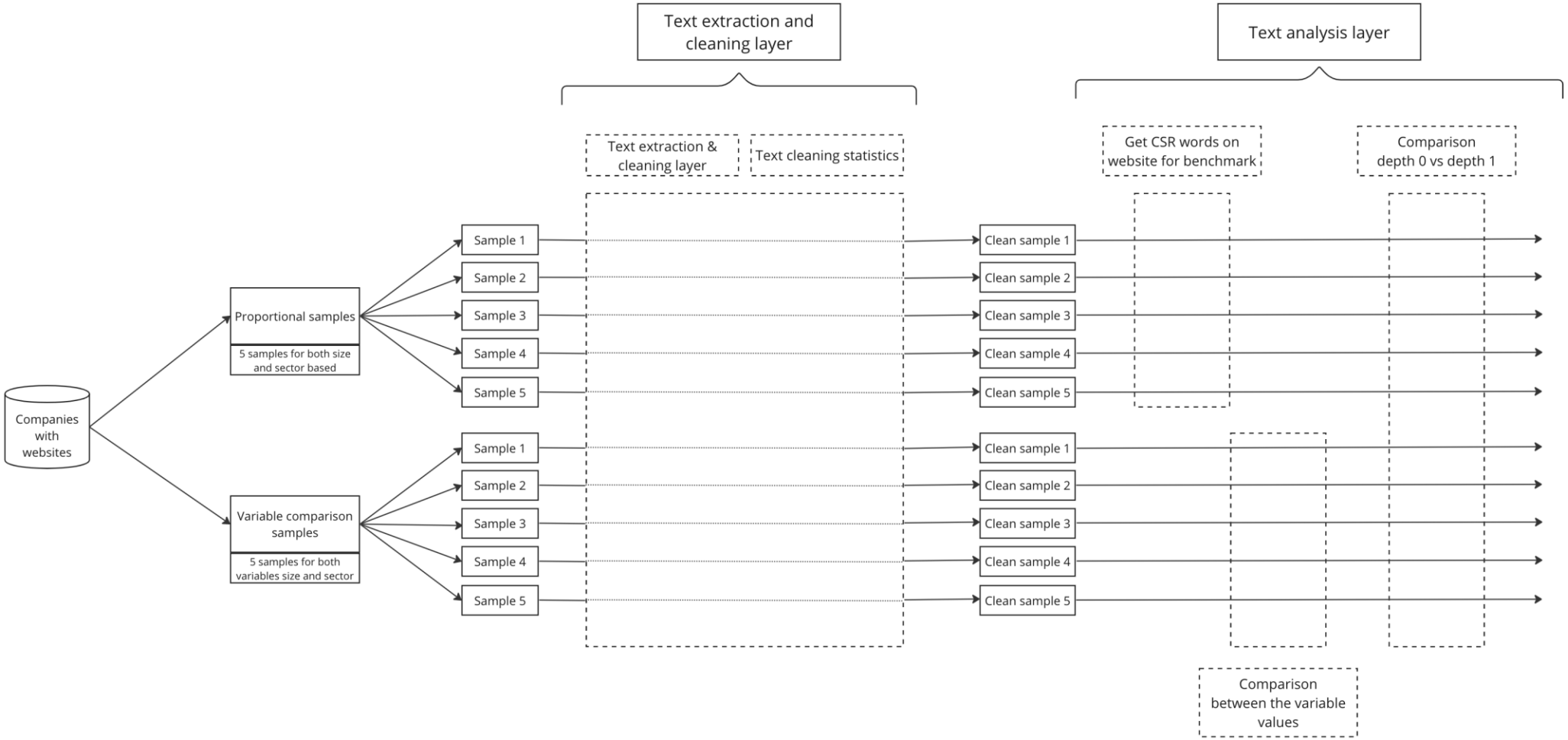
How can text extraction and natural language processing (NLP) techniques contribute in the **quantitative analysis of web scraped data** obtained from Belgian company websites, with the aim of **assessing the degree of corporate social responsibility (CSR)** commitment demonstrated by these organizations?

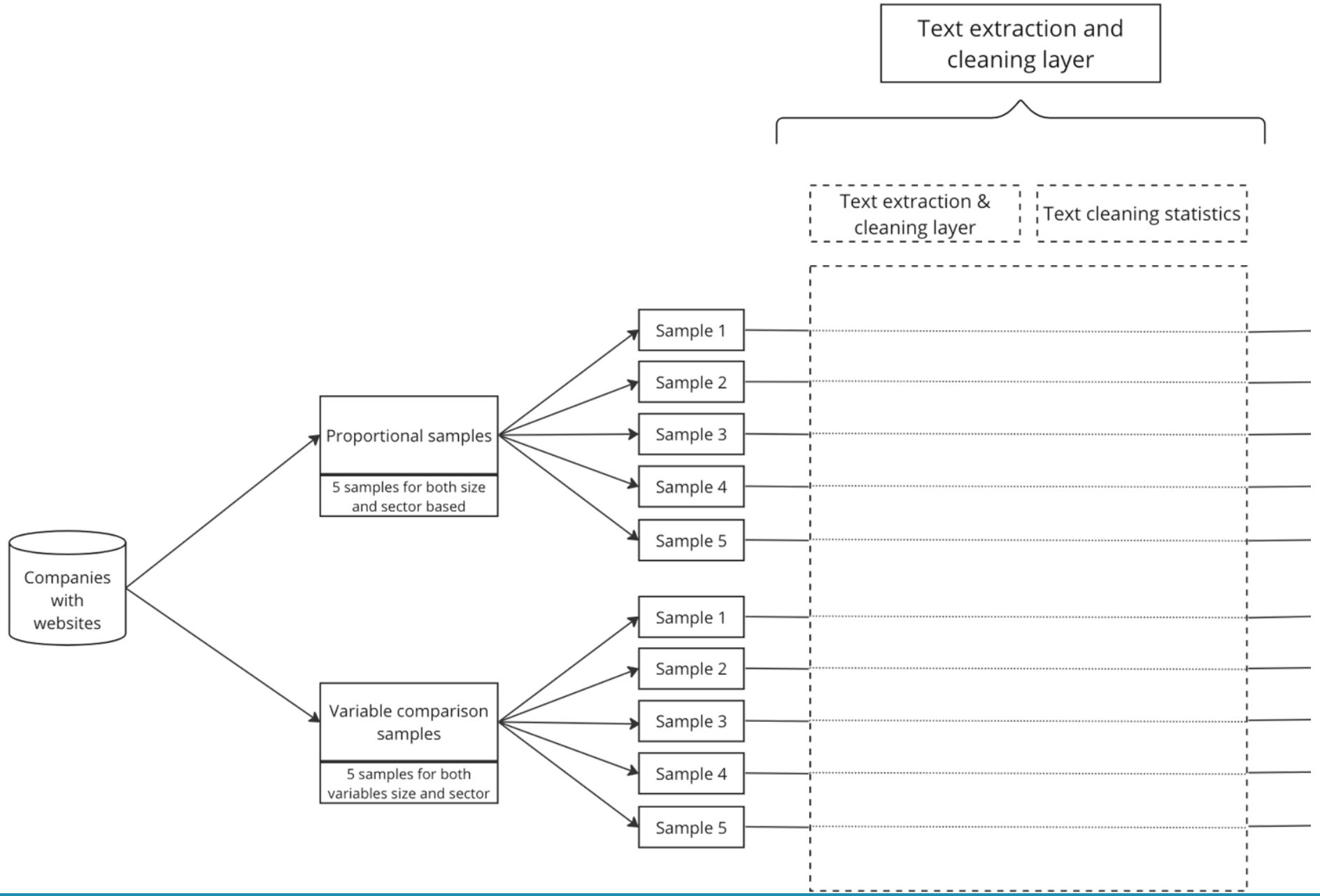
The side research questions are:

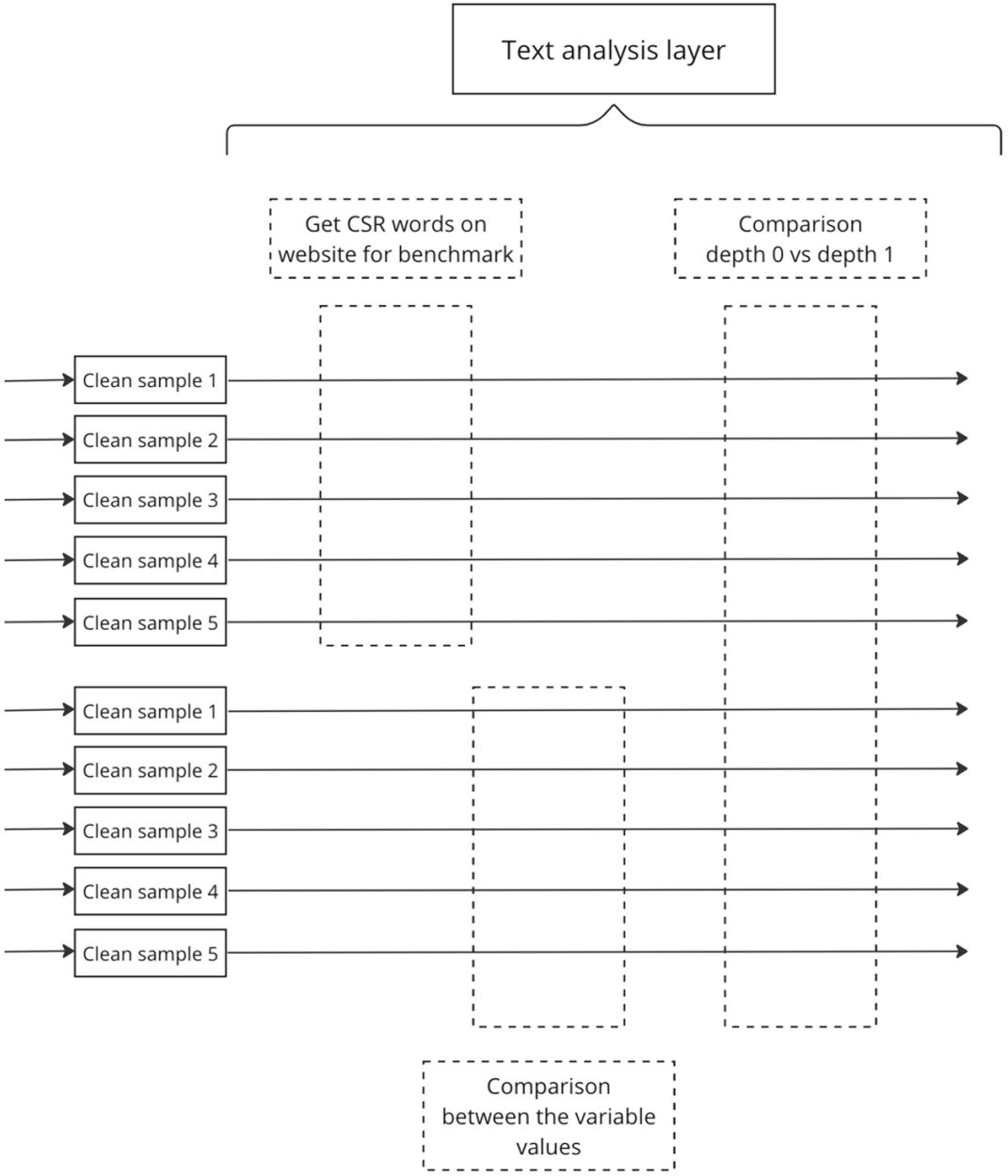
- Which **text pre processing steps** are the most effective in removing irrelevant data from the extracted text of company websites?
- What is the **correlation between different company characteristics and their displayed CSR commitment**? These characteristics include sector, size and whether they are listed on the stock exchange market.
- How can one **ensure the sustainability and generalizability of these techniques** to keep the research relevant and applicable to future trends?

Table of contents

Methodology visualisation



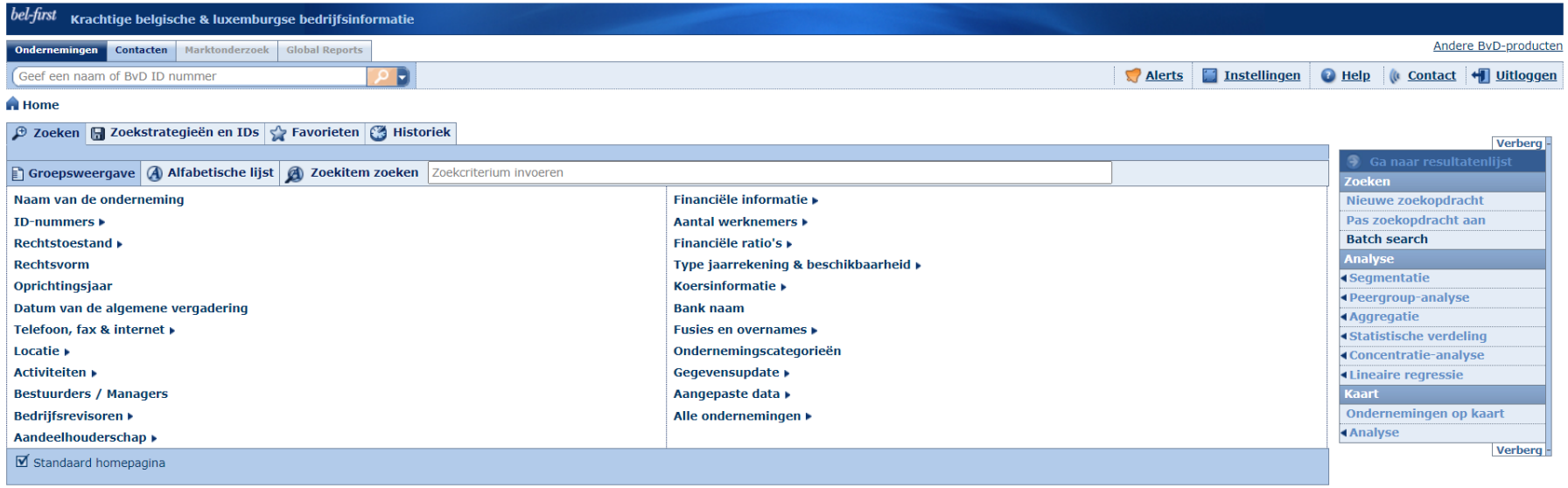

Text extraction and cleaning layer
Text extraction & cleaning layer
Text cleaning statistics
Companies with websites
Proportional samples
5 samples for both size and sector based
Variable comparison samples
5 samples for both variables size and sector
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5


Text analysis layer
Get CSR words on website for benchmark
Comparison depth 0 vs depth 1
Clean sample 1
Clean sample 2
Clean sample 3
Clean sample 4
Clean sample 5
Clean sample 1
Clean sample 2
Clean sample 3
Clean sample 4
Clean sample 5
Comparison between the variable values

Which companies to include?

- Web URL
- Company Size
- Company Sector (NACE-code)
- Other variables

BUREAU VAN DIJK
A Moody's Analytics Company



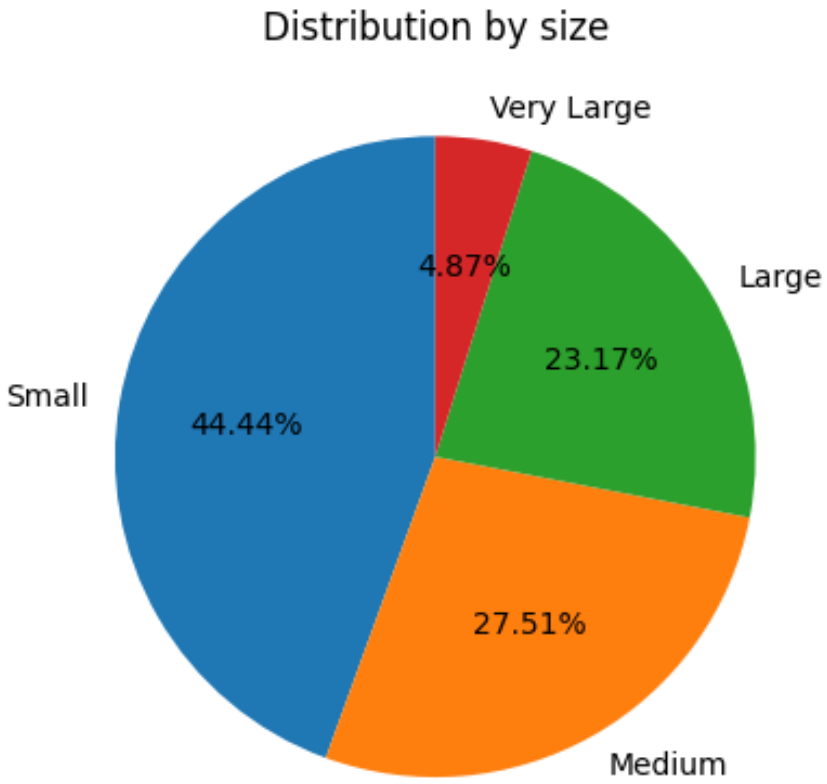
Size & sector distributions

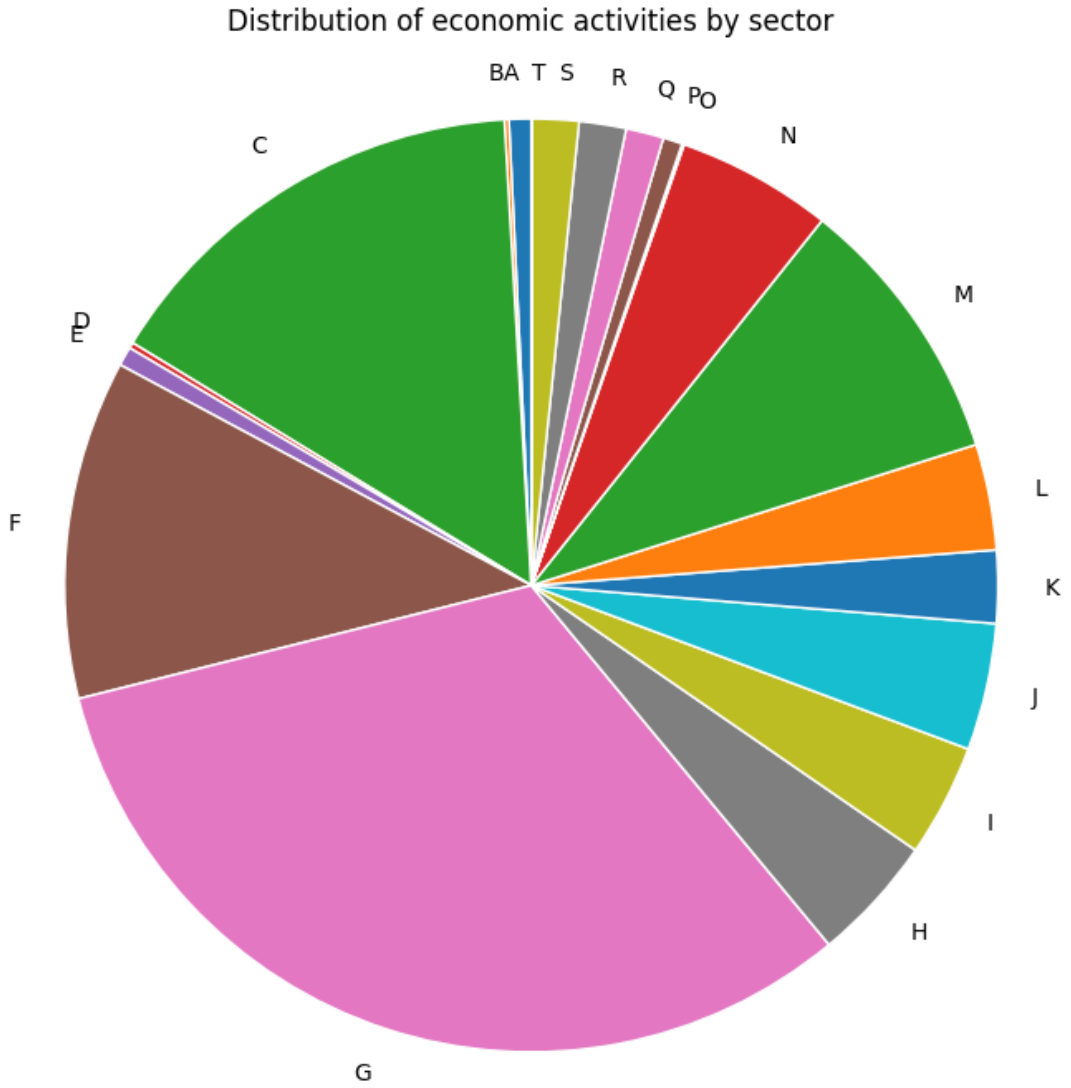
Size	Count	Percentage
Small	13485	44.44%
Medium	8348	27.51%
Large	7030	23.17%
Very Large	1478	4.88%

Table 1. Distribution by size

Letter	Description	Count	Percentage
A	Agriculture, forestry and fishing	231	0.76%
B	Mining and quarrying	48	0.16%
C	Manufacturing	4666	15.38%
D	Electricity, gas, steam and air conditioning supply	61	0.20%
E	Water supply; sewerage, waste management and remediation activities	201	0.66%
F	Construction	3563	11.74%
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	9743	32.10%
H	Transportation and storage	1330	4.38%
I	Accommodation and food service activities	1183	3.90%
J	Information and communication	1334	4.40%
K	Financial and insurance activities	765	2.52%
L	Real estate activities	1112	3.67%
M	Professional, scientific and technical activities	2855	9.41%
N	Administrative and support service activities	1637	5.40%
O	Public administration and defence; compulsory social security	23	0.08%
P	Education	202	0.67%
Q	Human health and social work activities	397	1.31%
R	Arts, entertainment and recreation	493	1.62%
S	Other service activities	487	1.61%
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	10	0.03%

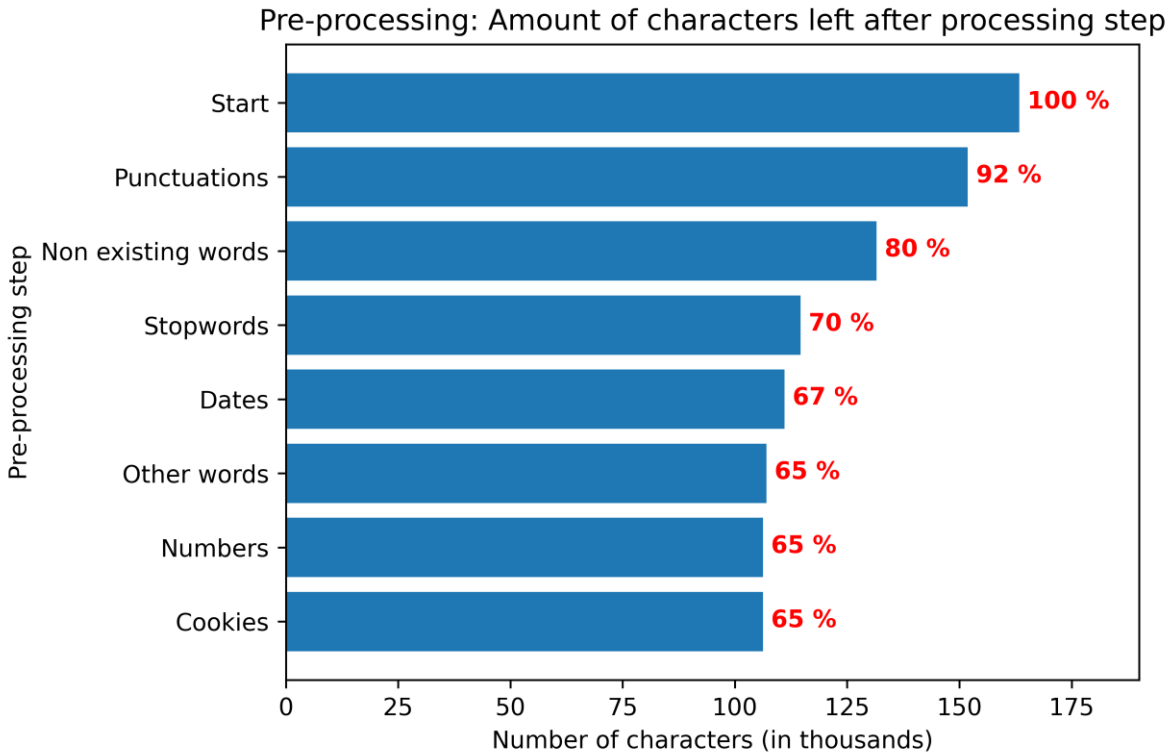
Table 2. Distribution of economic activities by sector

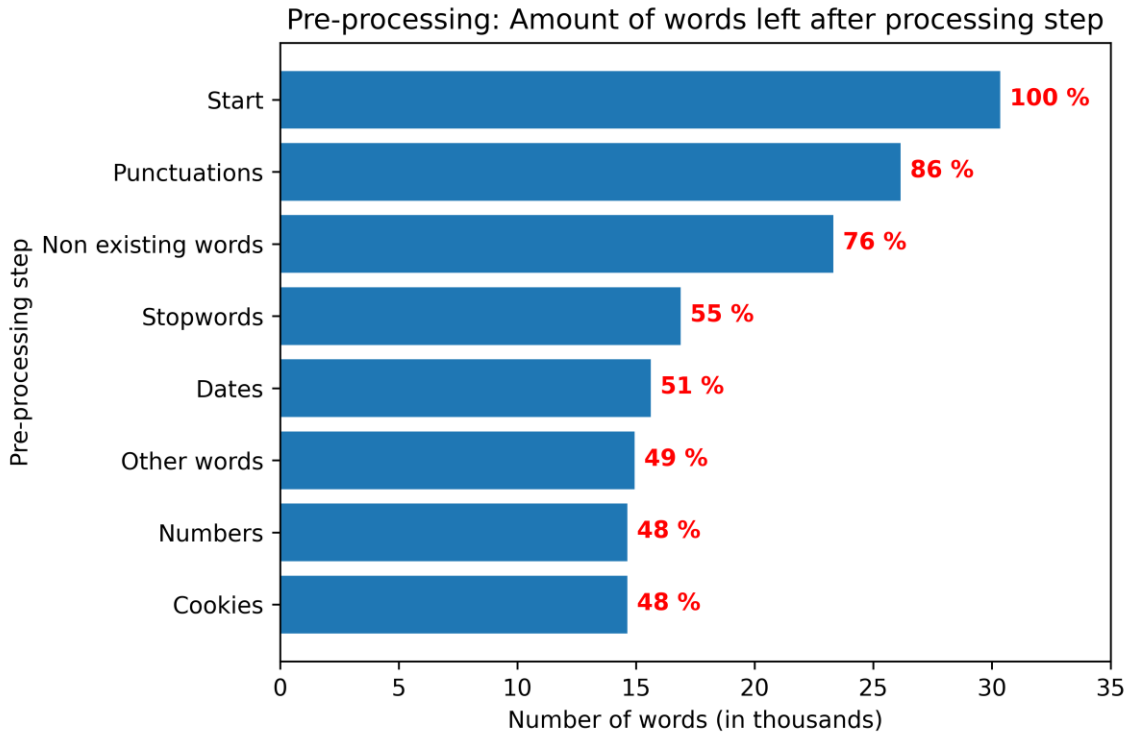




Company Size **sample 4** percentages:
Small: 57.50% (115 companies)
Medium: 24.50% (49 companies)
Large: 17.00% (34 companies)
Very Large: 1.00% (2 companies)

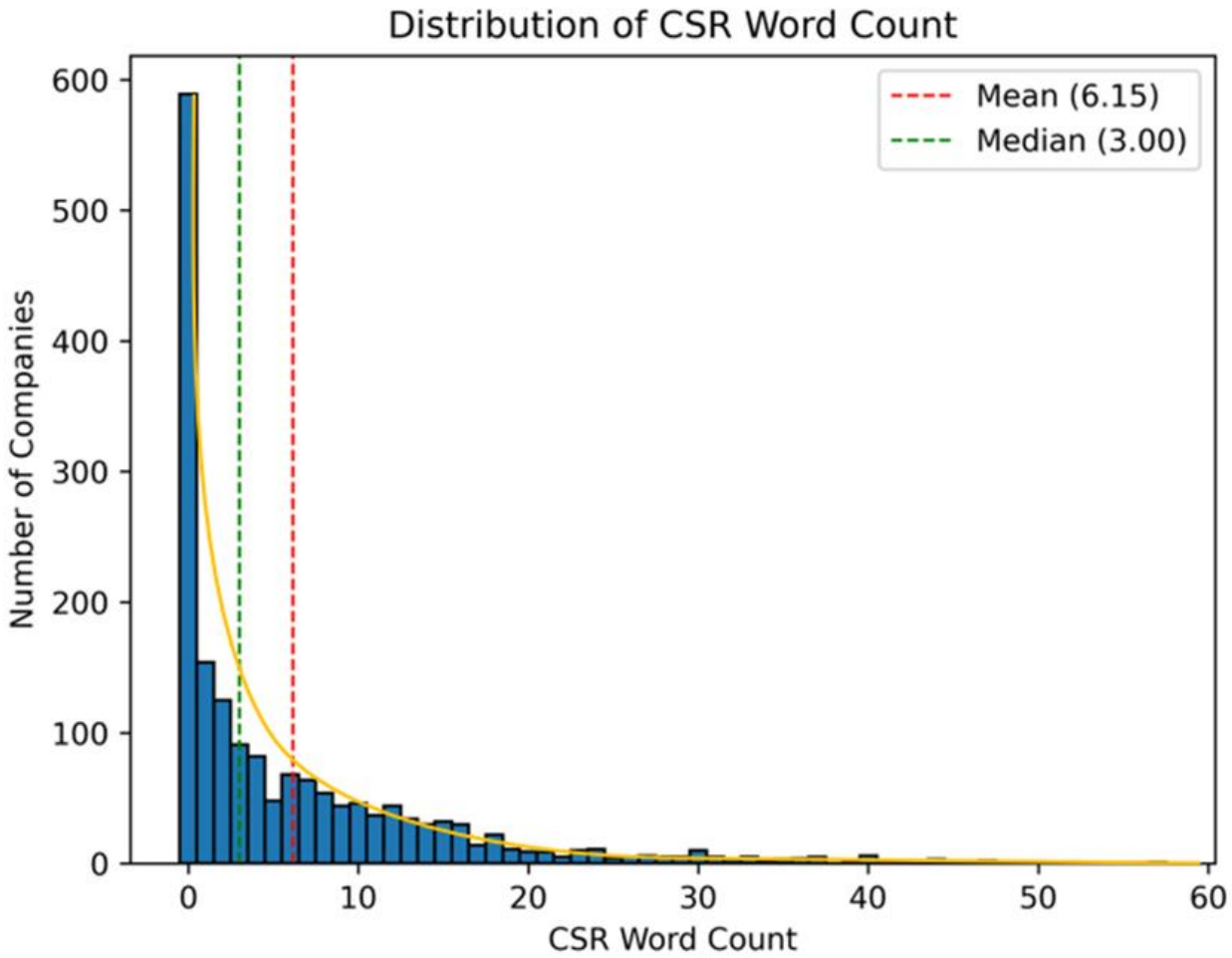
Pre-processing results





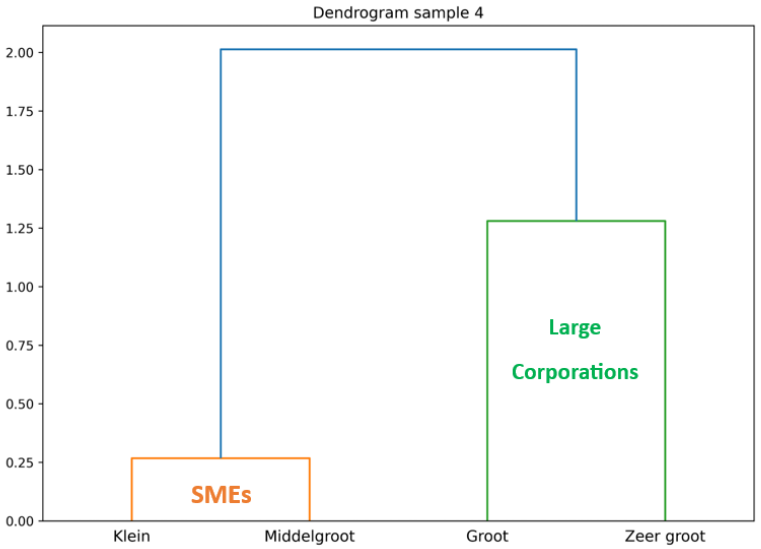
CSR communication of ‘the average Belgian company’

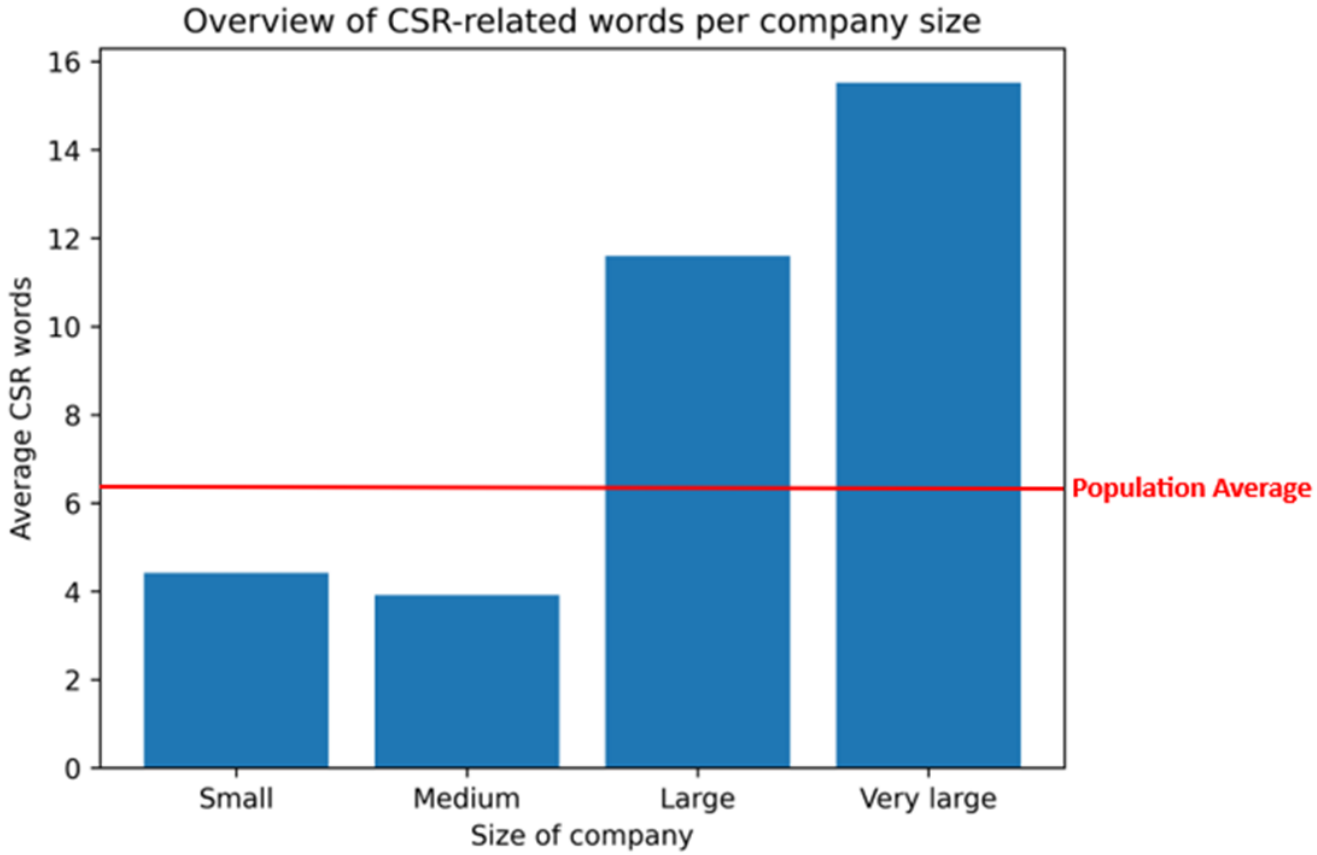
- 10 samples
- Benchmark
- Right skewed data
  - Mann-Whitney U tests



CSR communication results (Size)

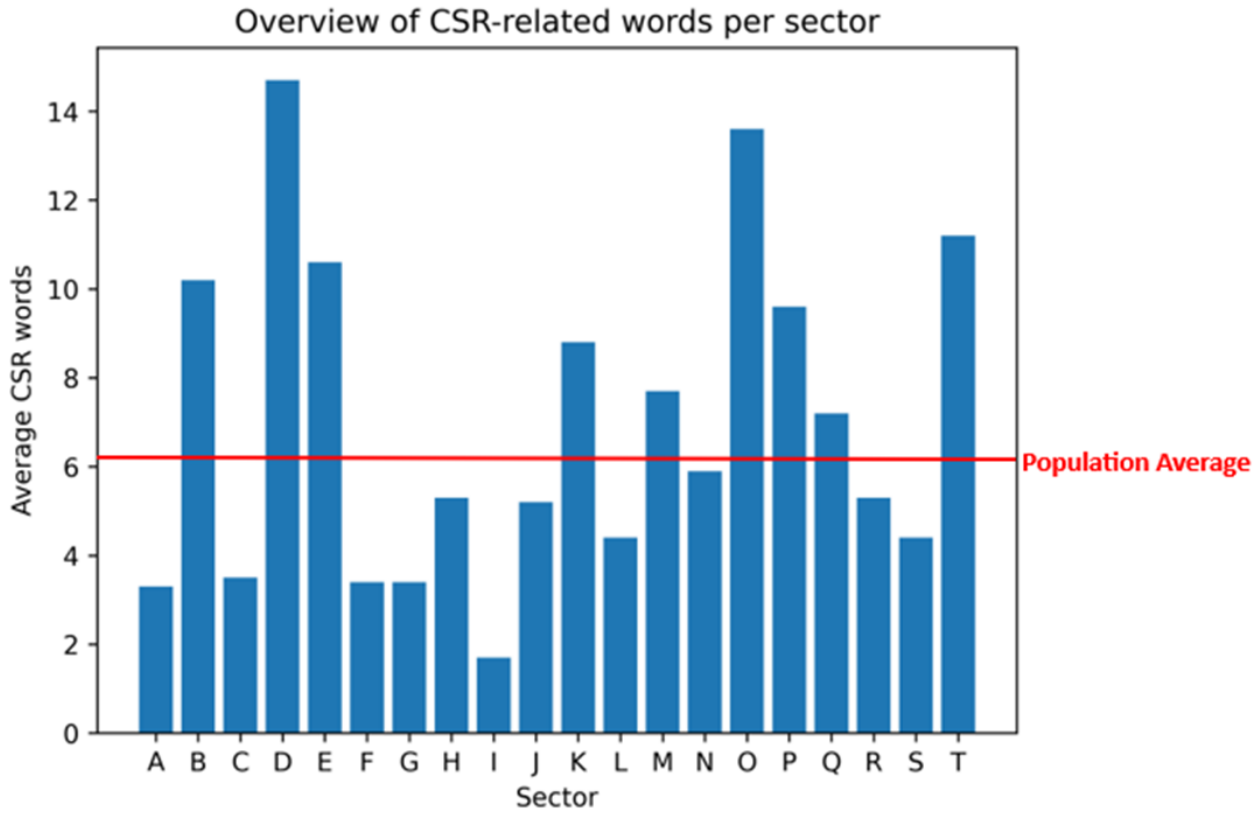
- Clustered
  - SMEs vs Large Corporations





CSR communication results (Sector)

- Compared to benchmark
- Only O & T significantly different 5 times
  - Public administration and defence; compulsory social security
  - Activities of households as employers; undifferentiated goods- and services-producing of households for own use



Biases + Future Research

Biases

- Sampling
- Selection
- Confirmation
- Publication
- Others
  - Website size
  - Blocked websites
  - Communication based

Future Research

- Generalizable
  - Other topics
- Practical testing of our results

Thank you for your attention, any questions?