PredictLeads

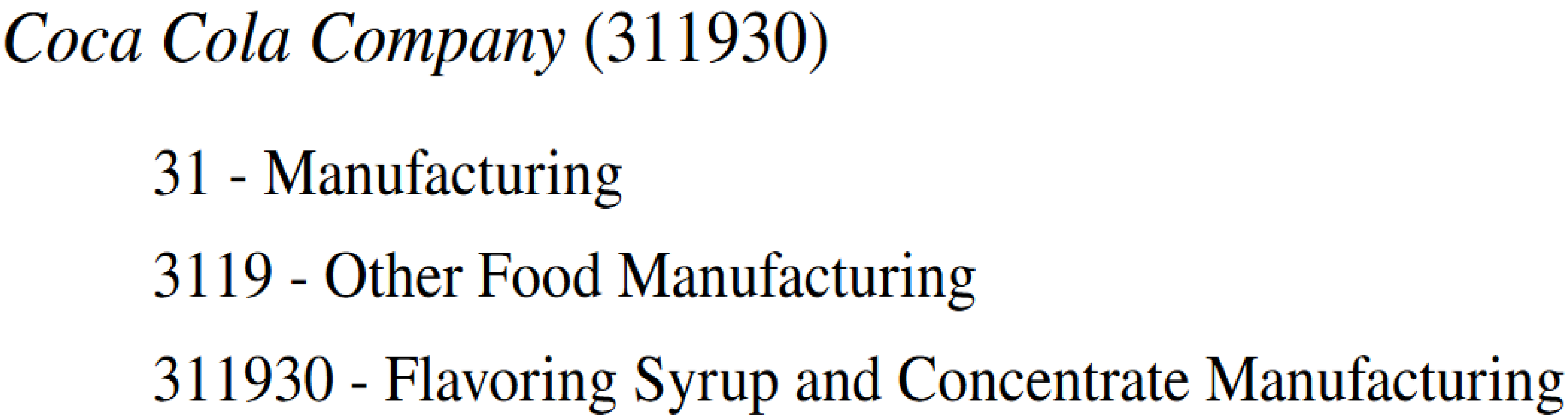
Jernej Avsec, Luka Bajić

Advisor: prof. Erik Štrumbelj

Classifying companies by industry

THE NAICS CODE

- North American Industry Classification System
- 1000+ distinct codes
- Goal: classify from company information



Coca Cola Company (311930)

31 - Manufacturing

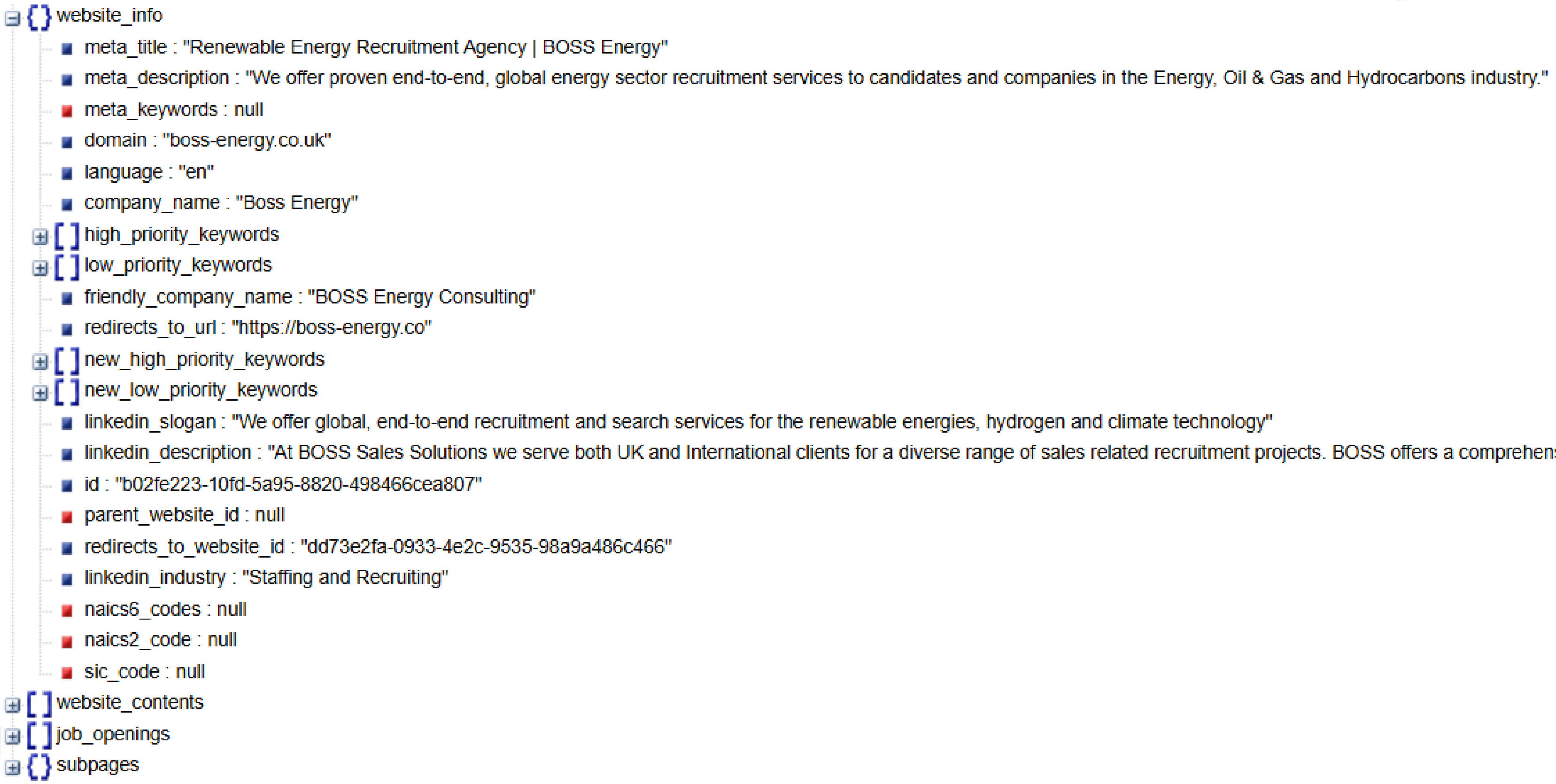
3119 - Other Food Manufacturing

311930 - Flavoring Syrup and Concentrate Manufacturing

Data

- PredictLeads provided textual data for 32k companies

- We additionally scraped official NAICS data



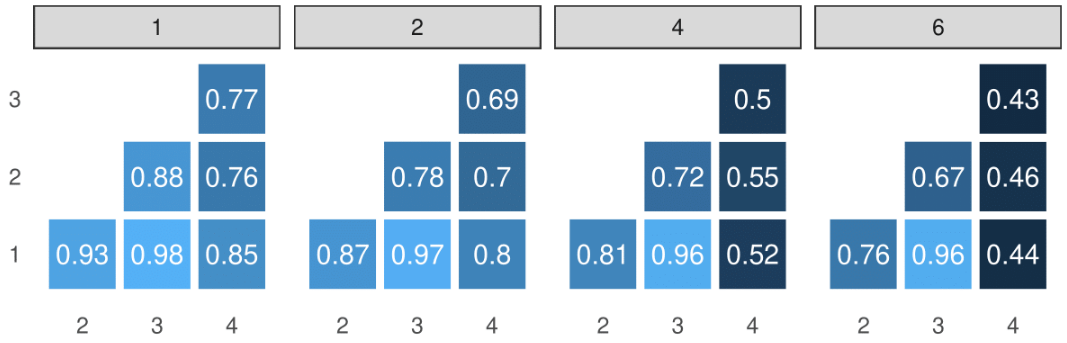
```
{ } website_info
    meta_title : "Renewable Energy Recruitment Agency | BOSS Energy"
    meta_description : "We offer proven end-to-end, global energy sector recruitment services to candidates and companies in the Energy, Oil & Gas and Hydrocarbons industry."
    meta_keywords : null
    domain : "boss-energy.co.uk"
    language : "en"
    company_name : "Boss Energy"
  [ ] high_priority_keywords
  [ ] low_priority_keywords
    friendly_company_name : "BOSS Energy Consulting"
    redirects_to_url : "https://boss-energy.co"
  [ ] new_high_priority_keywords
  [ ] new_low_priority_keywords
    linkedin_slogan : "We offer global, end-to-end recruitment and search services for the renewable energies, hydrogen and climate technology"
    linkedin_description : "At BOSS Sales Solutions we serve both UK and International clients for a diverse range of sales related recruitment projects. BOSS offers a comprehen
    id : "b02fe223-10fd-5a95-8820-498466cea807"
    parent_website_id : null
    redirects_to_website_id : "dd73e2fa-0933-4e2c-9535-98a9a486c466"
    linkedin_industry : "Staffing and Recruiting"
    naics6_codes : null
    naics2_code : null
    sic_code : null
[ ] website_contents
[ ] job_openings
{ } subpages
```

Ground truths

- Problem: outdated/invalid codes

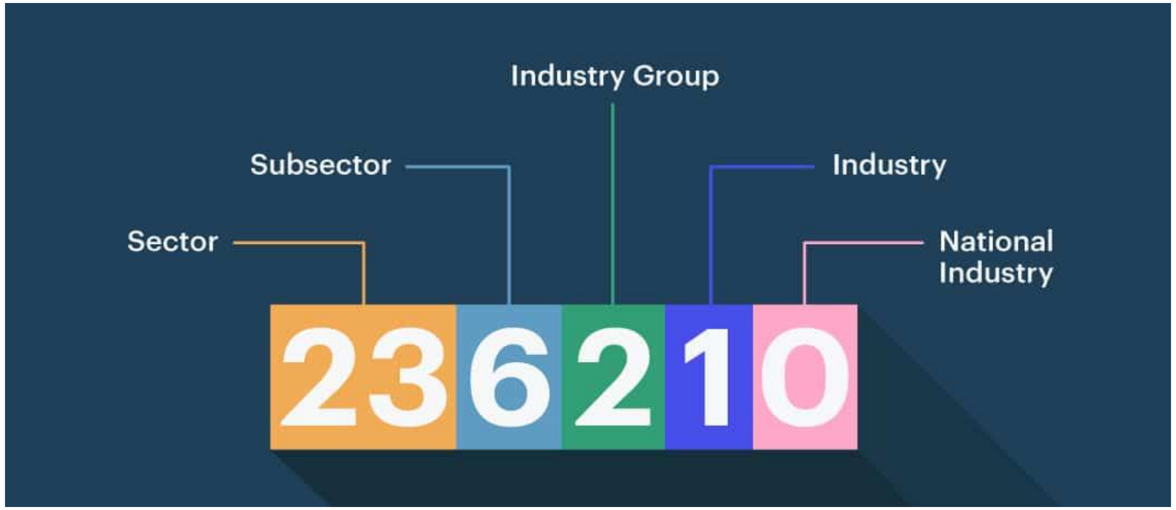
- 5 sources:
  - Official NAICS codes
  - 3 human annotators
  - LLM-generated ground truth

- Large disagreement between annotators

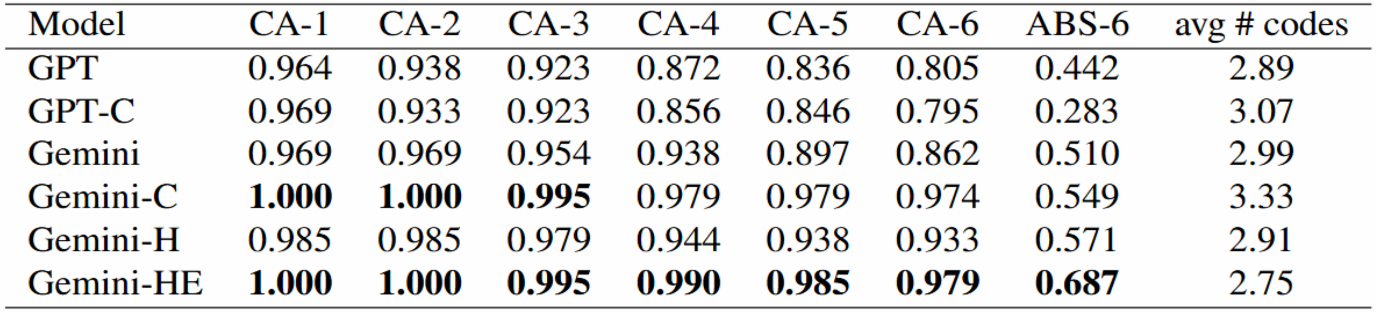


Metodology Part 1

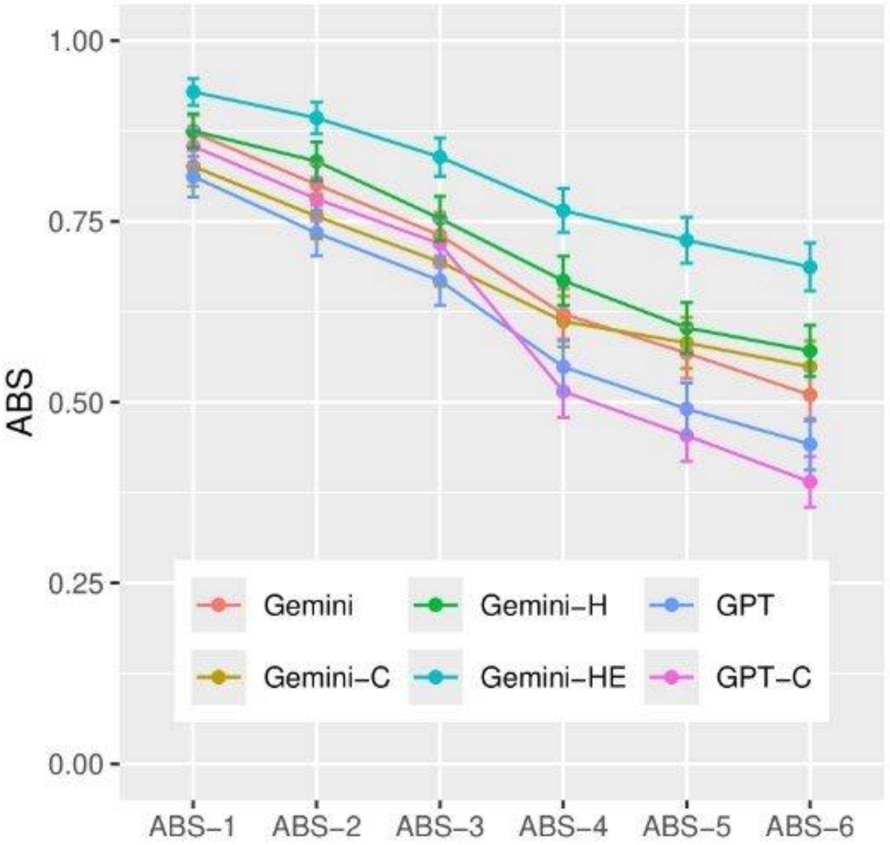
- Proof of concept from PredictLeads
- Different models, prompts
- Hierachical classification



Results

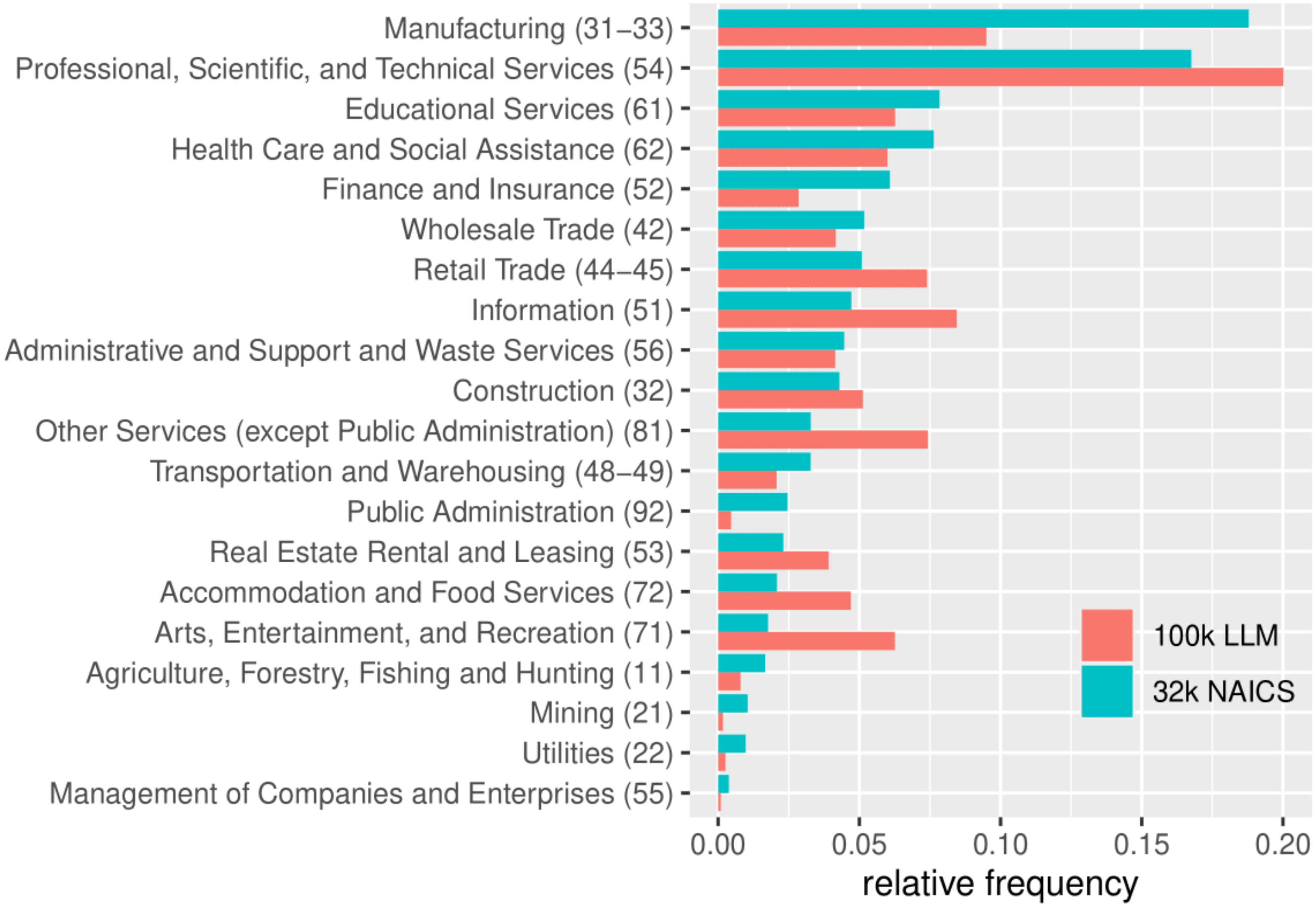


Model	CA-1	CA-2	CA-3	CA-4	CA-5	CA-6	ABS-6	avg # codes
GPT	0.964	0.938	0.923	0.872	0.836	0.805	0.442	2.89
GPT-C	0.969	0.933	0.923	0.856	0.846	0.795	0.283	3.07
Gemini	0.969	0.969	0.954	0.938	0.897	0.862	0.510	2.99
Gemini-C	**1.000**	**1.000**	**0.995**	0.979	0.979	0.974	0.549	3.33
Gemini-H	0.985	0.985	0.979	0.944	0.938	0.933	0.571	2.91
Gemini-HE	**1.000**	**1.000**	**0.995**	**0.990**	**0.985**	**0.979**	**0.687**	2.75



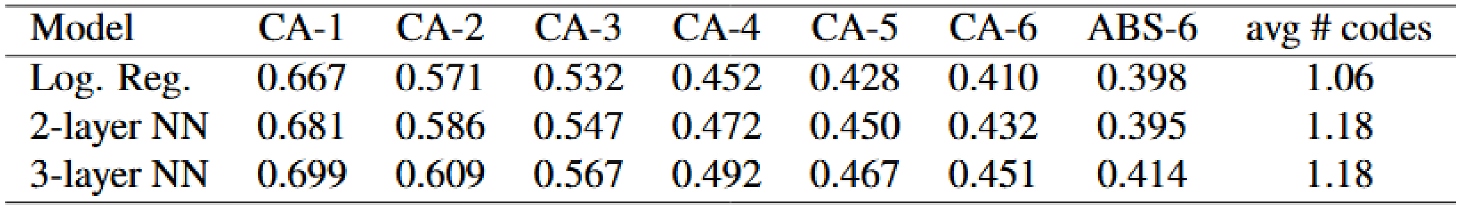
Second dataset

- Extremely well performing LLM
- ~230€ for 100,000 companies
- ~11 hours for 100,000 companies



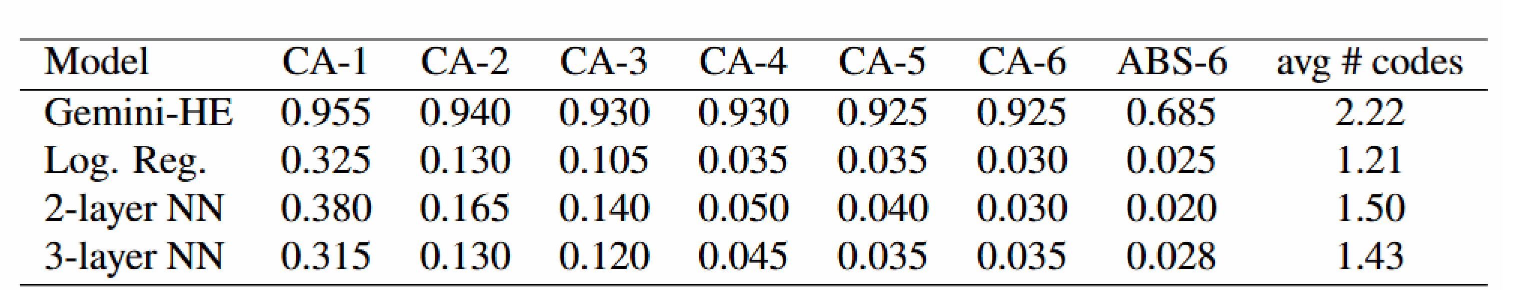
Methodology and results Part 2

- Using Gemini results as ground truth
- Embedded company metadata with ModernBERT
- Trained logistic regression and 2, **3 layer neural network.**



Model	CA-1	CA-2	CA-3	CA-4	CA-5	CA-6	ABS-6	avg # codes
Log. Reg.	0.667	0.571	0.532	0.452	0.428	0.410	0.398	1.06
2-layer NN	0.681	0.586	0.547	0.472	0.450	0.432	0.395	1.18
3-layer NN	0.699	0.609	0.567	0.492	0.467	0.451	0.414	1.18

Validation set



Model	CA-1	CA-2	CA-3	CA-4	CA-5	CA-6	ABS-6	avg # codes
Gemini-HE	0.955	0.940	0.930	0.930	0.925	0.925	0.685	2.22
Log. Reg.	0.325	0.130	0.105	0.035	0.035	0.030	0.025	1.21
2-layer NN	0.380	0.165	0.140	0.050	0.040	0.030	0.020	1.50
3-layer NN	0.315	0.130	0.120	0.045	0.035	0.035	0.028	1.43

Test set

Conclusion

- Operationally acceptable performance with LLM.
- Not good enough for scalability, fast enough.