Veridion Project

Company Classifier

Author: Andreea Ichim

2025

Introduction

This project serves as an application for the DeepTech Engineer Intern role at Veridion.
The goal of this project is to build a robust company classifier according to a new insurance taxonomy.

Datasets Description

For this project, I started with two datasets:

- One containing company details (description, business_tags, sector, category, niche).
- Another containing the insurance taxonomy.
-

The companies dataset

```
companies = pd.read_csv("data\ml_insurance_challenge.csv")
companies.head(3)
```

	description	business_tags	sector	category	niche
0	Welchcivils is a civil engineering and constru...	['Construction Services', 'Multi-utilities', '...	Services	Civil Engineering Services	Other Heavy and Civil Engineering Construction
1	Kyoto Vegetable Specialists Uekamo, also known...	['Wholesale', 'Dual-task Movement Products', '...	Manufacturing	Fruit & Vegetable - Markets & Stores	Frozen Fruit, Juice, and Vegetable Manufacturing
2	Loidholdhof Integrative Hofgemeinschaft is a c...	['Living Forms', 'Farm Cafe', 'Fresh Coffee', ...	Manufacturing	Farms & Agriculture Production	All Other Miscellaneous Crop Farming

The insurance taxonomy dataframe

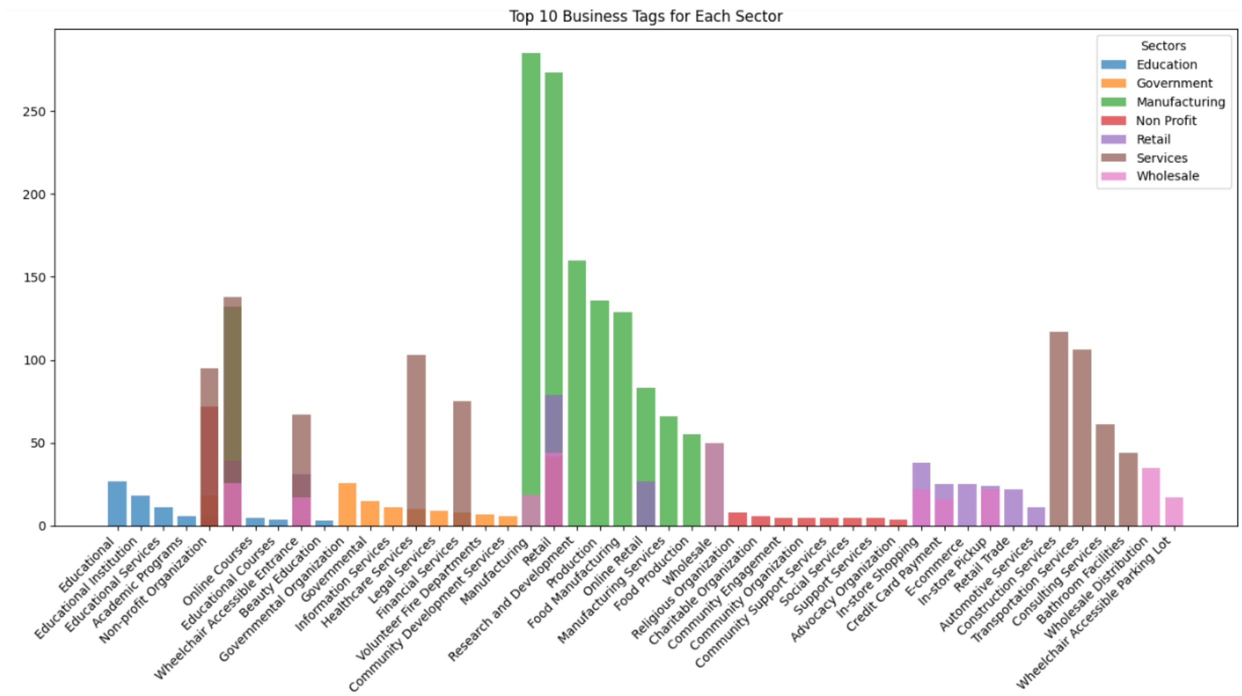
```
insurance = pd.read_csv("data\insurance_taxonomy.csv")
insurance.head(5)
```

	label
0	Agricultural Equipment Services
1	Soil Nutrient Application Services
2	Pesticide Application Services
3	Ornamental Plant Nurseries
4	Landscaping Services

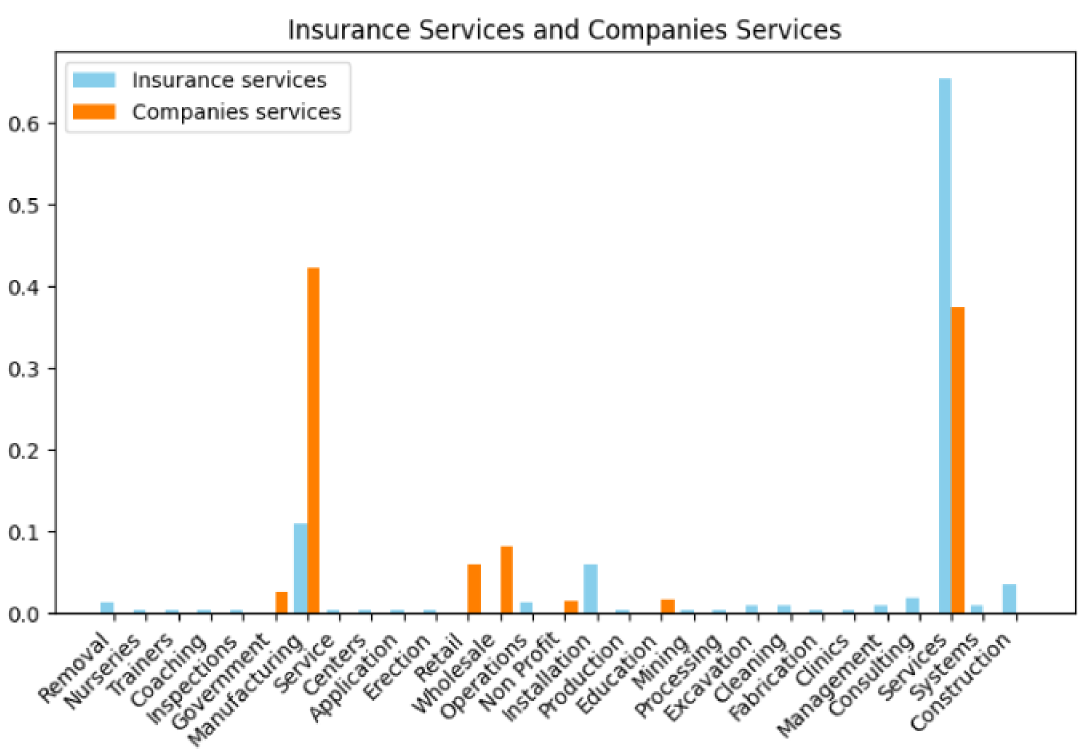
Data Analysis

At this stage, I examined if there were any null values in each column. Since the number of null values was very small compared to the total dataset, I decided to drop those rows.

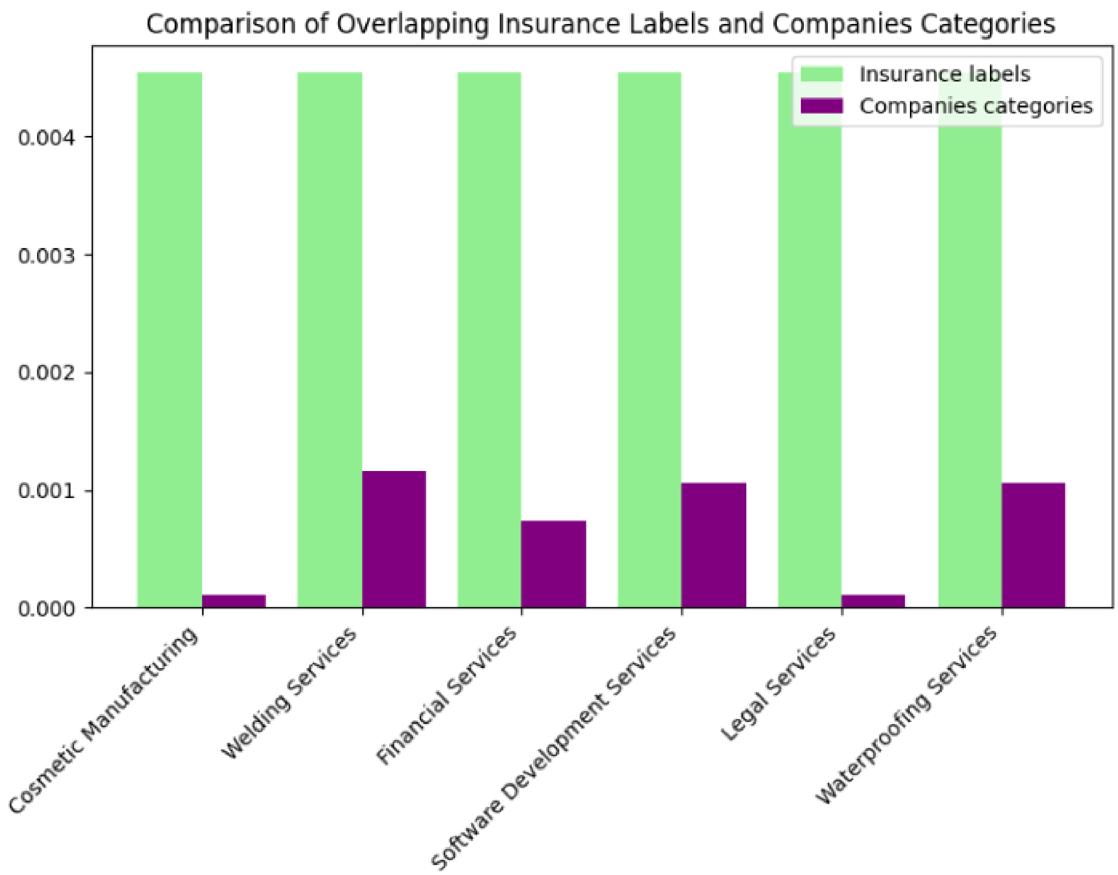
I separated all the business_tags into different columns to analyze their connection with the sector. The graph below shows that some business_tags are common within certain sectors.



Next, I tested my initial assumption that the last word in the taxonomy labels would match the sectors. However, this assumption was incorrect—only "Manufacturing" and "Services" were common between them.



I also explored the relationship between taxonomy labels and company categories, but I found only six similarities.



Formatting the Data for FastText

For FastText training, the data needs to follow a specific format*: __label__tag followed by a text describing something related to that tag.*

I decided to use the business_tags columns as labels and the rest of the columns as descriptive text, saving the final data into a final_text column.

```
companies["final_text"].iloc[0]
```

```
'__label__Construction_Services __label__Multi-utilities __label__Utility_Network_Connections_Design_and_Co
nstruction __label__Water_Connection_Installation __label__Multi-utility_Connections __label__Fiber_Optic_I
nstallation welchcivil civil engin construct compani special design build util network connect across uk of
fer multiutil solut combin electr ga water fibr optic instal singl contract design engin team capabl design
electr water ga network exist network connect point meter locat develop well project manag reinforc diver p
rovid custom connect solut take account ani exist asset maxim usag everi trench meet project deadlin welchc
ivil ha consid experti instal ga electr connect varieti market categori includ residenti commerci industri
project well civil engin servic heavi civil engin construct servic'
```

Splitting the Dataset

I split the dataset into:

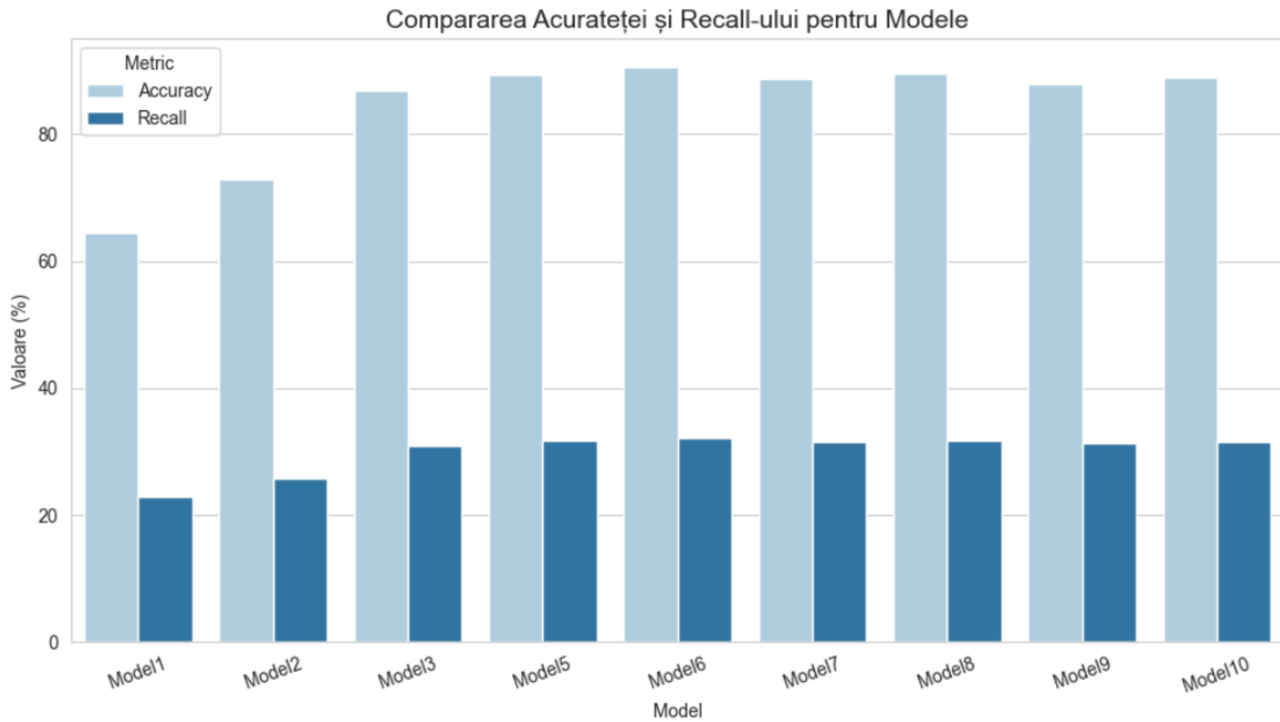
- **80% training set** and **20% test set**, stratified by companies["sector"].
- Then, I further split the training set into **90% training** and **10% validation**.

For training and validation, I saved only the final_text column as .txt files to train the model. For the test set, I kept the original columns and created a text column combining all fields except business_tags while normalizing the text.

Choosing a training model

To select a training model, I performed supervised training using FastText with different parameters. I then created a bar plot to compare the accuracy and recall of each model.

```
model = fasttext.train_supervised(
    input="data/test_data.txt",
    lr=0.1,
    epoch=5,
    wordNgrams=2,
    bucket=200000,
    dim=50,
    loss="ova"
)
```

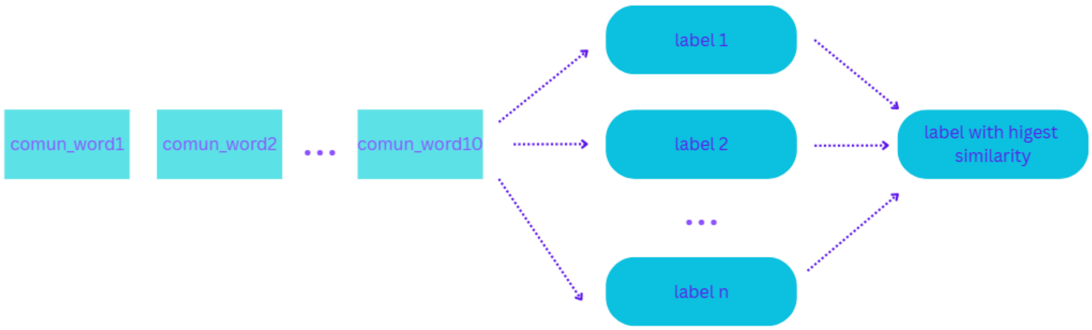


By testing different parameters, I aimed to improve recall. The highest values achieved were for **Model 6**, with:

- **Accuracy:** 90.49%
- **Recall:** 32.11%

Companies and Txonomy Comparation

In this step, I aimed to improve precision by selecting the top 10 most common words from final_text, vectorizing them, and comparing them to each vectorized row in the taxonomy's label_norm column.
The final insurance_label was assigned based on the label with the highest number of similarities.




```
companies_final = pd.read_csv("output7_results.csv")
companies_final.head(10)
```

	description	business_tags	sector	category	niche	insurance_label
0	Annakut Atta is an Australian company that spe...	Food Manufacturing, Distribution Network, Whea...	Manufacturing	Food Production	Flour Milling	Bakery Production Services
1	V Farms is an Australian company that speciali...	Pistachio Rootstock, Spring Planting, UCB1 Var...	Manufacturing	Farms & Agriculture Production	Peanut Farming	Agricultural Equipment Services
2	Clean Zero is a company that specializes in pr...	Water Surface Cleaning, Cleaning Products Manu...	Manufacturing	Cleaning Equipment & Supplies	Polish and Other Sanitation Good Manufacturing	Wood Product Manufacturing
3	Jardinerie Les Fleurs Bleues is an urban garde...	Graphic Foliage, Plant Decorations, Outdoor Fu...	Manufacturing	Plant Nurseries & Stores	Nursery and Tree Production	Ornamental Plant Nurseries
4	Hebei Yiwu Motor Manufacturing Co., Ltd. is a ...	Manufacturing, Largest Producer of Natural Gas...	Manufacturing	Stainless Steel Products	Iron and Steel Pipe and Tube Manufacturing fro...	Wood Product Manufacturing
5	More Grey Solutions Limited is a web solutions...	Web Solutions Provider, Drone Pilots, Security...	Services	Airline Companies	Nonscheduled Chartered Passenger Air Transport...	Crisis Management Services
6	Windy City Studios LLC is a decorative and fin...	Epoxy Resin Workshops, Stucco Marble Workshops...	Education	Fine Arts Schools	Fine Arts Schools	Arts Services
7	His & Hairs is a small family-run hair and bar...	Joico Select Salon Products, Hairstyling and B...	Services	Beauty Salons	Beauty Salons	Animal Day Care Services
8	Funderija Artistika Joseph Chetcuti is a found...	Bronze Restoration, Silicone Putty, Installati...	Manufacturing	Forging & Metal Stampings	Other Nonferrous Metal Foundries (except Die-C...	Sheet Metal Services
9	The company is involved in the production and ...	Plastic Extrusion Profiles Manufacturing, PVC ...	Manufacturing	Plastics Products	Plastics Material and Resin Manufacturing	Wood Product Manufacturing

Conclusions

Although FastText models achieved high accuracy, they struggled with recall. The best-performing model reached an accuracy of 90.49%, but its recall was only 32.11%, highlighting difficulties in classifying less frequent labels.

Moving forward, I can focus on improving recall by refining feature selection or exploring alternative embeddings beyond FastText. Additionally, experimenting with advanced NLP techniques, such as transformer-based models, could further enhance performance. I can also work on expanding taxonomy alignment by incorporating more contextual features from company descriptions.