

Sungjae Jin

Data Scientist / Data Analyst

Address	: 11 Mullet Road, North York, Ontario, M2M 2A7
Phone	: 437 – 818 – 7530
Email	: leopolt8th@gmail.com
GitHub	: https://github.com/leopolt8th-hub/Portfolio

Work Experience

- Deepnoid (Seoul, South Korea)
- AltheNutrigene (Seoul, South Korea)
- Misoinfotech (Seoul, South Korea)

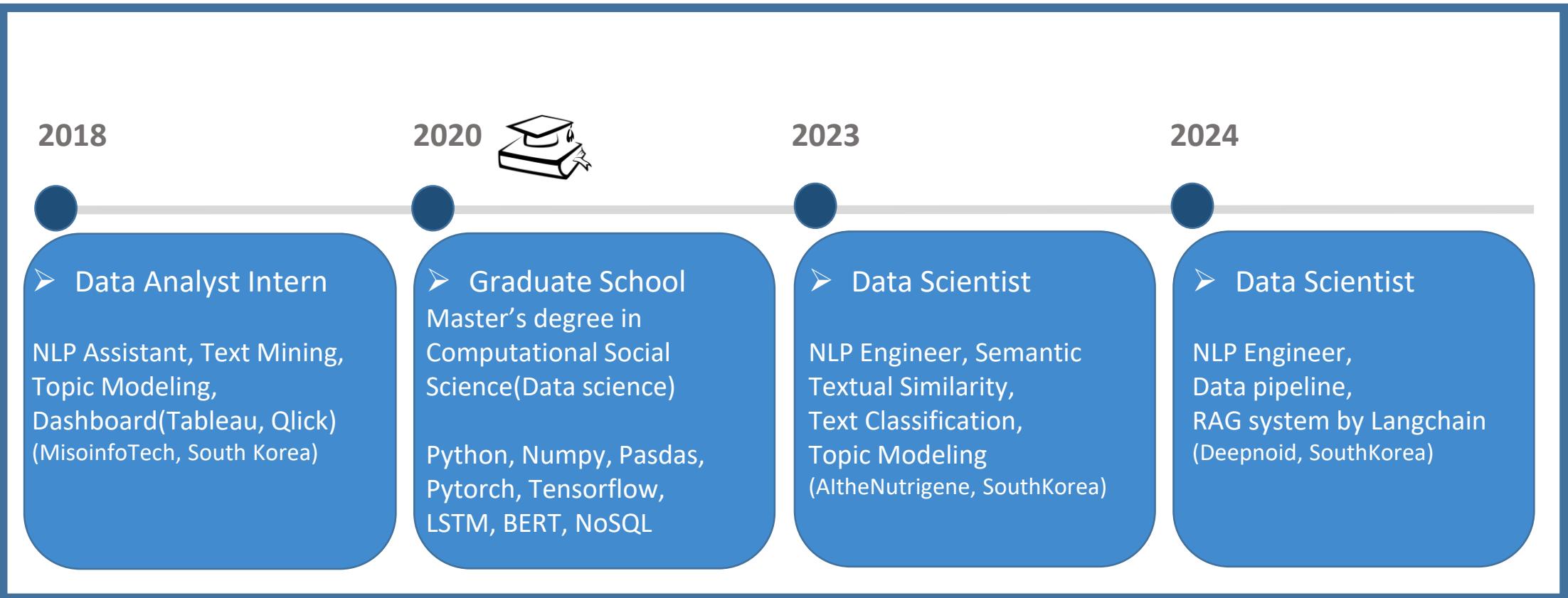
Education

- Hanyang University, Seoul, South Korea
- Master of Science Degree in Computational Social Science (2020-2023)
- Hanyang University, Seoul, South Korea
- Bachelor of Arts Degree in Sociology (2014-2020)

Skills

- Data Science (NLP, Python, Pytorch, Tensorflow, LSTM, BERT, RoBERTa, LLM fine-tuning, SQL, Spark, RAG)
- Data Analysis (Dashboard, Data Visualization, Statistic Analysis, Excel, PowerPoint, Microsoft Office, Tableau)

Time Line



Project 01.

Social data textual Classification

Project Title	A Study on the Social Factors of Low Birth rate from the Rational Choice Theory: Focusing on YouTube Comments
Institution / Company	Hanyang University, South Korea
Date	2021 ~ 2022

1) Project Summary

A study to find out what factors Koreans think are causing the low birth rate, resulting in the lowest birth rate in the world. From 33 YouTube videos of Korean public broadcasting related to low birth rate, 102,632 comments were crawled by Selenium. Crawled comments were classified by **LSTM**, **BERT**, and **KoBERT** models to 7 social factors and 5 emotions. Then it was analyzed by **Linear Regression analysis**, to see the relation between factors and emotions and analyzed correlation between 7 social factors. Natural language analyses were conducted through morpheme analysis and data visualization.

2) Problem Statement

1st problem was Crawling Social data of Youtube comments. Conducted adjusting Selenium Python library test by test.
And Tagging 7 social factors and 5 emotions were hard work. This work is done by 3 graduate students who majored in social science.
2nd problem was the multiple factors that the data had. Some of the comments contained multiple factors. It led to the correlation analysis between social factors. But to Linear Regression, needed independent variable. For that the emotion factors are tagged independently.

3) Dataset / Preprocessing

Data source : Youtube comments (33 YouTube videos of Korean public broadcasting related to low birth rate) crawled by Selenium
Dataset size : 102,632 -> 89,601 (delete noisy, irrelevant data)
Cleaning steps : Word extraction -> Tokenization -> labeling each comments (7 Social factors: jobs, housing costs, competitive culture, human instrumentation, stratification, national policy, and social environment. 5 emotions(despair, anger, ridicule, abandonment, and annihilation of comments)
Data imbalance : Handled using Data Augmentation of RD(random deletion), RS(random swap) for multi classification.

4) Methodology / Model

Methodology : From the Rational Choice Theory, To find which factors are causing the low birth rate to be the result of the world's lowest birth rate.
Model : BERT base, KoBERT, LSTM(Non BERT model for comparison)
Tools : Python, Pytorch, Scikit-learn, Pandas, Numpy, Matplot, Tokenization: Konlpy(for Korean language)

5) Evaluation

<Linear Regression>							
Category	Jobs	Housing costs	Human instrumentation	Competitive culture	Stratification	National policy	Social environment
Despair	0.0114	0.0453	0.1080	0.0225	0.0940	0.0781	0.9223
Anger	0.2419	0.1928	0.3280	0.2150	0.2938	0.3671	0.8344
Ridicule	0.0773	0.1013	0.3735	0.1897	0.5234	0.2003	0.7462
Abandonment	0.0662	0.0508	0.1936	0.0018	0.0539	0.0067	0.2515
Annihilation	0.3866	0.6062	0.6546	0.4672	0.7216	0.4903	0.8807

<7 factors Correlation>



<Word Cloud>



<Metrics>

Model	Accuracy	Precision	Recall	F1-score
LSTM	0.81	0.5	0.35	0.41
KoBERT	0.88	0.88	0.89	0.88
BERT	0.95	0.95	0.90	0.88

Project 02.

Financial Spending Category Classification

Project Title	Financial Spending Category Classification
Institution / Company	AitheNutrigene (Ubibelox Project)
Date	2023. 06. ~ 2023. 09.

1) Project Summary

This project is analyzing from customers' financial data which is withdrawn and paid expenditure details. For this, 2 preprocessing work is used. Account details data entered by customers or automatically entered are arbitrarily entered or non-standardized languages. As 1st step, This should be made to be recognized with the same meaning through standardization and dictionary work. 2ndly **NER(Named Entity Recognition)** is used. NER is the task of tagging words in a document according to a predefined category for important words(keywords or classifiable items). It is used to extract important information from documents, and the types of entity names classified include person name, institution name, location, date, time, currency, and number. The success of entity name recognition depends on contextual understanding, language diversity, and processing power of domain-specific vocabulary. 3rd from NER, use core keyword for ML to learn and fine-tune BERT models. Finally made BERT model is used for actual use.

2) Problem Statement

Data Preprocessing : Understanding domain knowledge in finance and customers' habits were the 1st struggle to handle data.

Data Pipeline : Building data pipeline from Data Base to language model to Dashboard needed collaborate with Backend coworkers.

Data imbalance : Handled using Data Augmentation of RD(random deletion), RS(random swap) for multi classification.

3) Dataset / Preprocessing

Data source : Customer "MyData" who agree to provide account expenditure information

Dataset size : 250,000(Numbers of expenditures by customers) for this project

Cleaning steps : Word extraction from NER -> Category keyword selection and build keyword dictionary -> Labeling each categories (25 Categories: Supermarket, Convenience store, Housing, Cafe/snack, Health, Savings, Insurance, Car, Tax, Education, Transportation, Hobbies/Leisure, Donations/Sponsors, Beauty, Restaurant, Subscription, Travel, Pet)

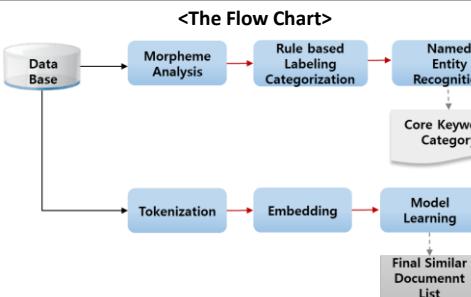
4) Methodology / Model

Methodology : Named Entity Recognition

Model : Fine tuned BERT, KoBERT, BiLSTM(Non BERT model for comparison)

Tools : Python, Pytorch, Scikit-learn, Pandas, Numpy, Matplotlib, Tokenization: Konlp(y(for Korean language)

5) Evaluation



<Metrics>

	precision	recall	f1-score
0.0	0.97	1.00	0.98
1.0	0.94	0.97	0.96
2.0	1.00	0.81	0.90
3.0	1.00	0.95	0.97
4.0	1.00	0.94	0.97
5.0	0.88	0.79	0.83
...			
accuracy			0.94
macro avg	0.92	0.93	0.92
weighted avg	0.95	0.94	0.94

Project 03.

Public Civil Complaint Analysis

Project Title	Public Civil Complaint Analysis for Seoul 3 district (KLID, Korea Local Information research & Development institute)
Institution / Company	AitheNutrigene (KLID Project)
Date	2023. 09. ~ 2023. 12.

1) Project Summary

A project to solve Civil Complaints quickly and efficiently in Seoul's autonomous districts with NLP technology. 1. standardize the categories and characteristics of Civil Complaints, 2. extract keywords by tokenization and proceed morpheme analysis. Then topics are categorized through topic modeling. 3. SentenceBERT is used to derive a list of high-similar previous Civil Complaints from the database by calculating the similarity between Civil Complaints.

2) Problem Statement

Data Preprocessing : Understanding domain knowledge in public and civil concepts.

Embedding selection : The selection of embedding methods was important to calculate sentence similarity.

3) Dataset / Preprocessing

Data source : Civil Complaints and Answers(already solved QA dataset)

Dataset size : 28,700 (Seoul's 3 autonomous districts Civil Complaint data)

Cleaning steps : Keyword extraction -> Tokenization -> Embedding

4) Methodology / Model

Methodology : Calculating Sentence similarity by Cosine Similarity

Model : SBERT Sentence Embedding for sentence level embeddings. SBERT trains BERT with Siamese networks to generate semantic similarity embeddings.

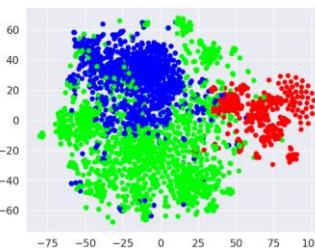
Tools : Python, Pytorch, Scikit-learn, Pandas, Numpy, Matplot, Tokenization: Konlpy(for Korean language)

5) Evaluation

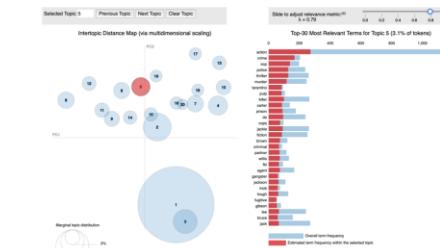
<The Flow Chart>



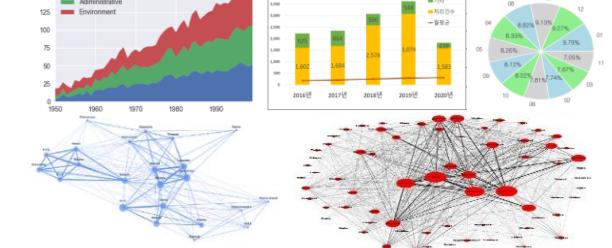
<t-SNE>



<LDA Topic Modeling>



<Keyword Analysis>





Thank you
For watching

Address

**11 Mullet Road, North York, Ontario,
M2M 2A7**

Phone

437 – 818 – 7530

Email

leopolt8th@gmail.com