PIER
PUEY UNGPHAKORN INSTITUTE
FOR ECONOMIC RESEARCH

Exploring the Thai Job Market Through the Lens of Natural Language Processing and Machine Learning

by

Nuttapol Lertmethaphat, Nuarpear Lekfuangfu, Pucktada Treeratpituk

January 2025
Discussion Paper
No. 228

Exploring the Thai Job Market Through the Lens of Natural Language Processing and Machine Learning

Nick Nuttapol Lertmethaphat (Bank of Thailand),

Warn Nuarpear Lekfuangfu (Universidad Carlos III de Madrid),*

Pucktada Treeratpituk (Puey Ungphakorn Institute for Economic Research),

August 2024

In recent decades, the Beveridge curve, which demonstrates a relationship between unemployment and vacancies, has emerged as a central organizing framework for understanding of labour markets – both for academic as well as central banks. The absence of consistent of the data in Thailand is a fundamental drawback in the utilisation of this important indicator. Data from online job platforms presents an alternative opportunity. However, the first and necessary step is to develop a process that can structure and standardise such data. In this paper, we develop an algorithm that standardise the high-frequency data from job websites, which consists of manually written job titles from major online job posting websites in Thailand (in Thai and English languages) into the International Standard Classification of Occupations codes (ISCO-2008), up to 4-digit level. With Natural Language Processing and machine learning techniques, our methodology automates the process to efficiently deal with the volume and velocity nature of the data. Our approach not only carves a new path for comprehending labour market trends, but also enhances the capacity for monitoring labour market behaviours with higher precision and timeliness. Most of all, it offers a pivotal shift towards leveraging real-time, rich online job postings.

Keywords: Labour market, Beveridge Curve, Online job platform, Machine Learning, Natural Language Processing, Text Classification, Thailand.

JEL Classifications: J2, J3, E24, N35

* Corresponding Author: Lekfuangfu, Department of Economics, Universidad Carlos III de Madrid, 126 Calle Madrid, Madrid, Spain, 28903, nlekfuan@eco.uc3m.es. Lekfuangfu acknowledges financial support from Puey Ungphakorn Institute for Economic Research (PIER research grant 2022/2).

1. INTRODUCTION

The Beveridge curve, which demonstrates a relation between unemployment and vacancies, has emerged in recent decades as a central organizing framework for understanding of labour markets – both for academic as well as central banks. The curve hinges upon the important role of vacancies in the determination of unemployment (Beveridge 1944, Elsby et. Al. 2015 for a review). The typically inverse relationship between job openings (vacancies) and job seekers (unemployed) has been shown to have fundamental implications for the efficiency of the matching process that generates employment relationships, and for the nature of shocks that drive fluctuations in the labour market. As a consequence, the Beveridge curve has played a pivotal role in debates over the functioning of labour markets, and has shaped the modern approach to understanding the coexistence and volatility of unemployment and vacancies.

Of the two components of the Beveridge curve, unemployment is relatively well understood as each country has regularly collected the statistics on a monthly basis with the Labour Force Survey – with the definition of unemployment follows universally what is set by the International Labour Organisation (ILO). With that, the time series of unemployment, at least at the national and sub-regional level, is commonly available for at least half a century.

In contrast, even among advanced economies, the statistics of vacancies and active jobseekers are less straight-forward. Therefore, only around the year 2000 that the reliable and timely data on job vacancies became available. In the US, the construction of the Beveridge curve relies on the Job Openings and Labor Turnover Survey (JOLTS).[1] Similarly in the UK, comprehensive estimates of job vacancies across the economy relies on the Vacancy Survey, which is a monthly survey that asks officially registered employers how many vacancies they are seeking outside their organization.

In the case of Thailand, to our knowledge, there is no consistent, timely, and regularly data on job vacancies or active jobseekers (administrative or survey). Therefore, it is not possible to construct a reliable measure of the Beveridge curve. In the past, the National Statistics Office and, at times, the Ministry of Labour commissioned the Surveys of Labour Demand. But the survey is not regular (as oppose to the monthly Labour Force Survey), and it covers only selected manufacturing establishments. This means, to date, policy decisions and academic debates have been executed in the absence of the Beveridge Curve.

One way to resolve this issue is to follow the footsteps of other economies and initiate a regular surveys of job vacancy at the national level. However, this approach may take some times as it requires strong commitments – human resources and financial investment – from multiple stakeholders.

An alternative is to make use of other data that is commonly available and informative of both key components of the curve. For the US, researchers and analysts of labour market also rely on the Conference Board's "Help-Wanted Advertising Index". Available since 1960, initial information that is used to construct the Help-Wanted Index came from job adverts in newspaper.

[1] JOLTS came into existence since 2000 (https://www.bls.gov/jlt/).

As job advertisement has now moved to online platforms, the current re-carnation of the index, HWOL, (by Conference Board and Burning Glass Technologies) comes from a large volume of online postings from multiple hosts.[2]

Unlike the survey data, job vacancies as well as resume posting truly reflect actual, real activities in the labour market. The data is also of a higher frequency (daily versus monthly) as well as more timely (real-time, without the delay of data processing). For this reason, a large and growing number of academic research now exploit the data of online job patterns (for instance, Hershbein and Kahn 2018, Deming and Kahn 2018, Hansen et. al. 2023). In the case of Thailand, Nakavachara and Lekfuangfu (2018) have already demonstrated potential applications of web-based data on job and resume posting. Using the data in one point in time, they calculated the ratio of number of vacancies to the number of resumes (v/u) at the national and regional levels. They are also able to show some degree of gender-bias patterns of both labour demand and labour supply among the firms and workers who conducted their job matching online. However, without more data over time, they are not able to construct a Beveridge curve. The online-based job data also has more potentials. First, with detailed information on both the employers (vacancy) and the workers (resume), one can further zoom into the labour market in more disaggregated dimensions – along the line of gender, age, sector, regional, education level, and even skill groups.

Above all, the information contained in the dataset grants a possibility to construct measurable components of a matching function of frictional labour market structure, namely posted wage, as well as reservation wage from the demand and the supply side of the labour market, respectively. To date, Thailand does not have such official data available. Even if researchers could imply the dynamic of labour supply in Thailand from well-known survey data sources (e.g., the Labour Force Surveys or the Socio-Economic Survey), the wage data from those surveys or the administrative data of the Social Security Office are those at the equilibrium, i.e., once the demand is matched with the supply. Therefore, labour-market related research has been limited by the absence of such information.

However, given the volume of the data (in the size of the observations as well as the detail in each observation), the execution is challenging and it requires the advancement of machine-learning technique (databasing, classification, natural language processing) to tackle and simplify the data. Considering that such data is not originally constructed for academic purposes and therefore is not documented in such classifications familiar to researchers, the first necessary step is to develop a process that can structure and standardise the data.

In this paper, we develop an algorithm that standardise the high-frequency data from job websites, which consists of manually written job titles from major online job posting websites in Thailand (in Thai and English languages) into the International Standard Classification of Occupations (ISCO) codes (up to 4-digit level). Through the integration of advanced Natural Language Processing (NLP) and machine learning techniques, our methodology automates what would otherwise be prohibitively labour-intensive due to the volume and velocity nature of the data. This enables us to tap into a vast, previously unexplored dataset, unlocking insights into

[2] See https://www.conference-board.org/topics/help-wanted-online for more details.

labour market dynamics with timeliness and depth not achievable through traditional survey data alone. This innovative approach not only carves a new path for comprehending labour market trends, but also enhances the capacity for monitoring labour demand with an unprecedented level of precision and timeliness. Our methodology marks a pivotal shift from reliance on delayed and less detailed survey data to leveraging real-time, rich online job postings.

In detail, our Machine Learning process is anchored in a dual-source dataset, with Thailand's Department of Employment (DOE) providing the primary data foundation as our training data. The DOE dataset encompasses hundreds of thousands of manually written records, offering a raw, unfiltered glimpse into the labour market's complexity of job posts, similar to those found on online job portals. In contrast, the ISCO codebooks (of 1988 and 2008 versions) serve as a complementary source, offering a structured, dictionary-like catalogue of ISCO codes paired with formal examples of occupations. While holding the standard of precise definition, the codebook presents a starkly formal language that often misses the nuanced, informal expressions typically found in actual job listings.

Leveraging the DOE dataset allows our model to adapt to the variability and complexity inherent in real-world job postings, mirroring the challenges we face when analysing data taken from online job portals. Simultaneously, the structured classifications provided by the ISCO codebook serve as a definitive standard. This dual-source strategy, combined with rigorous training, validating, and fine-tuning, aims to enhance the model's capability to deal with the vast array of data encountered in online job postings and resumes, equipping it to offer more accurate, meaningful insights into the labour market dynamics.

The strength of our methodology comes from the integration of advanced Natural Language Processing (NLP) and machine learning techniques to classify a vast amount of job titles from online portals into standardised ISCO codes. By exploring the application of two cutting-edge sentence embedders: the Universal Sentence Encoder (USE) and WangchanBERTa (developed by Thailand Artificial Intelligence Research Institute)[3] in converting unstructured job titles into structured numerical formats, we ensure semantic nuances are captured accurately. A novel cleaning process employing 'majority rule' and 'close-match' approaches is also utilised to tackle inconsistency and ambiguity inherent in our training data.

The core of the methodology lies in the evaluation of several candidate models, aimed at optimizing the classification task. The research delved into a rich spectrum of models, including the Cosine Similarity Classifier, which leverages the geometric intricacies of vector space, and the Hierarchical Cosine Similarity Classifier, designed to refine classification across ISCO's hierarchical structure. The exploration extended to ensemble learning techniques through the Random Forest and eXtreme Gradient Boosting (XGBoost) models, renowned for their robustness in processing complex datasets. Neural Networks were employed to capture deep, contextual representations of job titles, while BERTopic was utilized for its novel approach to topic modelling, providing additional layers of analysis. Each model's unique approach aims to navigate the complex task of classifying a vast array of job titles into precise ISCO codes.

[3] See https://airesearch.in.th/releases/wangchanberta-pre-trained-thai-language-model for more information.

Furthermore, every model is subjected to rigorous validation using the 5-fold cross-validation technique, ensuring a comprehensive assessment of their performance and suitability for the task at hand. Given the nature of the vast range of categories (over 400 groups of 4-digit ISCO codes) involved, accuracy was chosen as the main metric for evaluating the performance of these models due to its straightforward features.

Our exploration led us to identify an optimal combination of USE and the eXtreme Gradient Boosting (XGBoost) model, which outperformed other configurations, achieving an impressive classification accuracy rate close to 90%, as particularly effective in handling the nuances of this task. By integrating these advanced tools, we have dealt with the significant challenge of classifying a vast array of job titles into precise ISCO codes. Most of all, we can efficiently and accurately pursue further in-depth analysis of Thailand's labour market with this valuable dataset and other related ones.

2. INTEGRATING NATURAL LANGUAGE PROCESSING (NLP) INTO MACHINE LEARNING (ML)

2.1. Data Requirement

Classification algorithms, whether applied to textual or other types of data, fundamentally require annotated data to enable supervised learning algorithms to recognize patterns and establish connections between features and their corresponding classes. While access to colossal, well-structured, and squeaky-clean textual data is ideal, real-world situations rarely provide those. Nevertheless, at the bare minimum, two qualities of annotated data necessitated for this kind of execution are **(a) volume** and **(b) variation**:

(a) Volume

A paramount dataset is pivotal for achieving statistical significance, allowing the model to learn from a wide array of instances and effectively reflect the complexities of languages. Especially when embedding is crucially required in the NLP field, our selected classification algorithms will face high-dimensional input vectors. An insufficient amount of data could hinder the degree of freedom of the model, hence impeding its capability to maximize statistical interference and generalizations. This abundance in data quantity is necessary not only for training but also for the critical phases of model validation and testing. During these three stages, a comprehensive dataset ensures that the model is rigorously evaluated and fine-tuned, enhancing its performance and reliability in real-world usage.

In addition, the substantial volume of data also plays a crucial role in combating one of the major challenges in machine learning: overfitting, where a model is too closely tailored to the training data. This could be mitigated by exposing the model to diverse scenarios within a large dataset. Such exposures will help ensure the model's adaptability and robustness.

Fortunately, the NLP field has evolved rapidly and yielded several publicly available resources to aid such an insufficiency. Building a language model upon pre-trained models drastically helps with the performance of text classification. Focusing on text classifications with Twitter data, Nguyen et al. (2021)[4] concluded that pre-trained models could overcome big data requirements. They also emphasized that a good choice of text representation has more impact than adding more data.[5]

(b) Variation

The variation present within annotated samples plays a vital role in shaping the effectiveness of text classification models. Increased variation enhances the model's capacity to recognize the pattern between an extensive spectrum of writing styles and subtle nuances and their corresponding labels. Moreover, the diverse representation of sector-specific jargon, acronyms, and contextual variations enriches the model's adaptability across a diverse range of input data, enhancing its versatility and adaptiveness in handling unforeseen data scenarios once it has been deployed.

Moreover, balanced data distributions across different classes within the dataset is vital. Sufficient class representation ensures that the model receives enough exposure to diverse categories, mitigating the risk of bias towards overrepresented classes. Furthermore, it enhances its capacity for generalization. In cases where imbalanced classes are presented, researchers might consider resorting to upsampling or oversampling techniques, depending on the context and nature of the datasets. The achieved balance, even though sometimes manufactured, incubates a healthier learning process, enabling the model to recognize distinctions among different classes better, thereby improving its ability to classify unseen instances accurately. This will ultimately contribute to its overall performance and reliability in real-world applications.

On the contrary, some papers found the effect of variations to be minimal. Chai et al. (2013)[6] built text classification models with the purpose of automatically identifying health information technology (HIT) incidents in the FDA's MAUDE database. They found that the classification performance was similar on "balanced" (50% HIT) and "stratified" (0.297% HIT) datasets, with F1 scores of 0.954 for the stratified dataset and 0.995 for the balanced dataset.

2.2. Sentence Embedding

Converting text into a format understandable by machine learning models is highly crucial for enabling algorithms to process, interpret, and derive insights from textual data. It is considered to be the pillar of Natural Language Processing (NLP). Textual data, which is inherently unstructured, requires conversion into structured numerical formats for computational analysis. The evolution of text representation techniques has been essential in bridging this gap between

[4] See Nguyen, T. H., Nguyen, H. H., Ahmadi, Z., Hoang, T.-A., & Doan, T.-N. (2021) for details.
[5] In their cases, the BERTweet embedder is a perfect fit since it was pre-trained with short Twitter text, compared to BERT, which was trained on long Wikipedia articles
[6] See Chai et. al. (2013) for details.

human language and machine understanding, facilitating the development of complex NLP models.

Like other branches of data science, text-to-numerical techniques have exponentially evolved in the past few decades. Initially, methods like Term Frequency (TF) and Inverse Document Frequency (TF-IDF) were prevalent. TF computes the frequency of a word within a document and structures the data into a tabular-like format. The underlying idea appeared easily comprehensible and executable; however, it possibly exhibited an oversimplified nature. To mitigate that, TF-IDF was developed to minimize the drawbacks of TF by weighing the importance of a word in a document relative to its occurrence in a corpus, aspiring to emphasize the importance of key contexts while eroding the general ones.

The aforementioned techniques provide basic numerical representations but lack semantic understanding and context. The representation generated by TF-IDF, for instance, considers word frequency, but does not capture the functions of the words or their semantic meanings. In recent years, the invention of word and sentence embeddings, such as Word2Vec, GloVe, and FastText,[7] introduces the conversion of textual data to numerical vectors, which captures semantic relationships between words based on their contexts using neural networks. Unlike earlier methods, word and sentence embeddings facilitate the capture of semantic similarities, analogies, and context in a way that traditional approaches like TF and TF-IDF were unable to achieve.

In this paper, we highlight two state-of-the-art embedding large-language models (LLMs): **WangchanBERTa** and **Universal Sentence Encoder (USE)**. Each embedding model encapsulates unique strengths and advantages, qualifying them as viable candidates for our project.

WangchanBERTa

WangchanBERTa[8] is an advanced language model created and developed by the Thailand Artificial Intelligence Research Institute. Its underlying technologies and architecture were fostered with the principle of and even named after two widely known LLMs: (1) "BERT" (Bidirectional Encoder Representations from Transformers), and (2) "RoBERTa".[9] While inheriting the core architecture of BERT and RoBERTa, WangchanBERTa stands out due to its language-specific training corpus dedicated to Thai, a language that has been underrepresented in the NLP domain and used to be a huge struggle for NLP enthusiasts. This customization enables

[7] See Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013), Pennington, J., Socher, R., & Manning, C. (2014), and Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017) for the details for each model, respectively.
[8] See He, J., Liu, D., Zhang, C., & Xu, H. (2021).
[9] BERT was released in 2018 by Google Research, introduced the Masked Language Model (MLM) objective, where it masked random words in input sentences and tasked the model with predicting those masked words, and its allegedly more competent successor. RoBERTa was later released in 2019 by Facebook AI, encompassing the refinement of BERT's approach, including further optimizations in training methodology, such as dynamic masking strategies and eliminating the next sentence prediction (NSP) task.

WangchanBERTa to better capture the nuances, syntax, and semantics specific to the Thai language, which illustrates a potential fit for our project.[10]

True to its specific objective, this model claims to have an exceptional ability to comprehend the nuances of the Thai language and deliver accurate predictions across diverse NLP applications, including token classification, sentiment analysis, part-of-speech tagging (POS), named-entity recognition (NER), and certainly, sentence embedding.

Universal Sentence Encoder (USE)

The Universal Sentence Encoder (USE) was built upon earlier advancements in word embeddings like word2vec but aimed for higher achievements: sentence-level embeddings.[11] Following its first release by Google Research in 2018, USE has grown to be a state-of-the-art language model equipped with capabilities for understanding and encoding textual information into fixed-dimensional vectors, providing versatile representations for sentences in 16 languages, including Thai.[12] Throughout its development, USE has been evaluated across numerous transfer tasks, such as sentiment analysis, subjectivity classification, question classification, semantic textual similarity, and more. These evaluations consistently emphasize the superiority of sentence-level embedding over word-level embedding. Furthermore, its versatile embedding capabilities have been empirically proven through a wide number of applications in the NLP field.

2.3. Our Dataset

The majority part of our dataset is comprised of the extensive job posting archive from Thailand's Department of Employment (DOE). This governmental institution serves as a mediator

[10] The iterative training process for WangchanBERTa employed a vast corpus of Thai language data, encompassing approximately 78.5 GB worth of text sourced from diverse origins such as news articles, Wikipedia, social media texts, and web-scraped content. Enriched with a colossal amount of data, the model underwent an extensively iterative training regimen, which took approximately 4 months to reach 500,000 training steps of the MLM training. By masking certain tokens in input sentences and training the model to predict the masked tokens based on the surrounding context, this training technique enhances the model's ability to understand and generate coherent text by learning the relationships between words within sentences. Thus, words can be encoded neatly into vectors regarding their locations in the sentence, variation of meanings, and overall context due to the model's vast and accurate comprehension of semantic meanings.

[11] See Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017).

[12] The training of the USE models involves a combination of unsupervised and supervised data sources. Unsupervised data is drawn from various web sources, including Wikipedia, web news, question-answer pages, and forums. This is supplemented with supervised training using the Stanford Natural Language Inference (SNLI) corpus. The design of the model consists of two primary encoding models: one based on the transformer architecture and another on the Deep Averaging Network (DAN). Even though they were designed to perform similar tasks, the transformer-based architecture prioritizes high accuracy in encoding sentences, even at the expense of greater model complexity and increased computational resources. On the other hand, the DAN architecture aimed for more efficient inference, sacrificing a bit of accuracy for faster computation. This allows the model to capture diverse language patterns and semantics as well as increase the versatilities of the transfer learning performance across diverse NLP tasks while calibrating the trade-offs as per users' idiosyncratic preferences

for both employers posting jobs and job seekers submitting resumes. With its nationwide reach through a network of regional offices, the DOE aggregates a substantial volume of job postings from various sectors, including those from remote areas of the country. This database is particularly valuable because it provides a focused snapshot of the demand side of the job market, capturing the diversity and scale of employer needs across the country. The breadth and depth of the DoE's dataset make it an optimal primary source for training our model, ensuring robustness in handling the complex variations typical of real-world job market data.

To enhance the accuracy of our classifications, we also utilize the ISCO-08 handbook as a complementary data source. Although smaller in scope compared to the DOE's database, the ISCO-08 provides a structured, formal description of occupations, which is essential for ensuring our model's alignment with correctly termed internationally recognized standards. The ISCO-08 framework offers a detailed breakdown of occupations into 415 unit groups [13], arranged hierarchically into minor, sub-major, and major groups based on skill level and specialization.

The combination of these two diverse data sources allows our classification models to be both versatile and adept for real-world application. The DOE's database introduces realistic variation in job titles, mimicking the informal and varied styles found in online job postings. This variability is crucial for training our model to recognize and process the wide range of expressions and terminologies used across different job platforms. Simultaneously, the formal and precise language of the ISCO handbook helps standardize the model's output, ensuring that the classified job titles align with internationally recognized occupational standards. By leveraging these strengths, our approach is positioned to effectively tackle the challenges of classifying online job postings, enhancing the reliability, accuracy, and versatility of our model across different job post platforms, thereby making it a robust tool for enhancing labour market analysis.

2.4. Embedding Models

The classification of web-scraped job titles into corresponding the ISCO, commonly used in economics and national survey data, presents a unique challenge in NLP. Therefore, to ascertain which embedders most effectively capture and interpret the nuanced lexicon of the labour market, we present an empirical evaluation of two state-of-the-art embedding models, WangchanBERTa and the Universal Sentence Encoder, within a structured test framework tailored to measure their linguistic precision and efficacy in both Thai and translated English.

2.4.1. The Three-Stage Test Framework

The efficacy of the embedding technologies is evaluated through a meticulously crafted three-stage test framework. This framework is specifically designed to assess the technical

[13] The major groups 0 and 6 were excluded from the study as they belong in unique industries where hiring process rarely appears in online job posts.

capability of the embedders across varying degrees of occupational similarity, each stage heightens the challenge of linguistic ambiguity and contextual overlap.

Stage One: Distinct Occupational Fields

The first stage serves as the proving ground for the embedders' basic functionality. The objective is to differentiate occupations that are distinct in their roles, specialties, and terminologies. For example, between 'Mechanical Engineers' (ISCO: 2144) and 'Building Architects' (ISCO: 2161). This test is anticipated to be a relatively straightforward task for any sophisticated embedder, providing a baseline for their performance.

Stage Two: Same Field, Different Roles

Building on the foundational assessment, the second stage introduces increased complexity. It examines the embedders' ability to distinguish between occupation titles that fall in the same broad field, but instead, carry nuanced differences in roles and specialties. For example, 'Generalist Medical Practitioners (Physicians)' (ISCO: 2211) and 'Dentists' (ISCO: 2261). The embedders must navigate through common medical terminology to prove their capabilities in accurately identifying the unique aspects of each job title.

Stage Three: Closely Similar Titles, Different on Skill Level Spectrum

The pinnacle and the most serious challenge of the testing framework is the third stage, which simulates a real-world scenario where the distinction between job titles is subtle yet significant in our context. For example, we task the embedders with differentiating between 'Accountants' (ISCO: 2411), 'Accounting Associate Professionals' (ISCO: 3313), and 'Accounting and Bookkeeping Clerks' (ISCO: 4311). On the one hand, these three job titles share thematic similarities. On the other hand, they exhibit distinct variations in professional hierarchy and skill levels. This stage evaluates the embedders' precision in capturing the nuanced gradience and skill differentiations inherent in closely related occupations.

2.4.2. The Process

Each test stage begins with the selection of two job title pairs, each pair representing the same ISCO code. These titles, originally in Thai, undergo translation into English. The translation serves a strategic purpose, which is to determine the most compatible language for each embedding technology. The goal is to identify the language that, when paired with the embedder, exhibits the ability to capture the nuanced terminologies of the labour market in vector representation and yields the highest accuracy, facilitating more effective subsequent application in supervised machine learning models, which we will discuss extensively in the upcoming sections.

To rigorously evaluate the embedders, we deploy *cosine similarity* as a measure of semantic proximity in high-dimensional vector space.[14] High cosine similarity values near 1 are indicative of vectors with similar directionality, suggesting thematic and semantic cohesion—a desirable characteristic we anticipate seeing among job title samples drawn from the same ISCO category. Conversely, low cosine similarity values denote vectors with divergent orientations, an attribute essential for distinguishing between different ISCO categories, critical for effective classification, and thus the metric we choose to determine the superior embedder for our further analysis.

2.4.3. Determining the Optimal Embedder for Accurate ISCO Code Classification

The results from stage one, as illustrated in Figure 1, recalibrate our understanding of the presumed capabilities of WangchanBERTa and USE across linguistic boundaries.

[14] This metric, symbolized by cos(θ), computes the cosine of the angle θ between two non-zero vectors, providing a scale from -1 through 1 to represent the degree of similarity, where 1 indicates identical directionality, and -1 represents completely opposite vectors. Formally, cosine similarity is defined by $\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$ where A and B denote the sentence embedding vectors for comparison. See more at https://studymachinelearning.com/cosine-similarity-text-similarity-metric/.

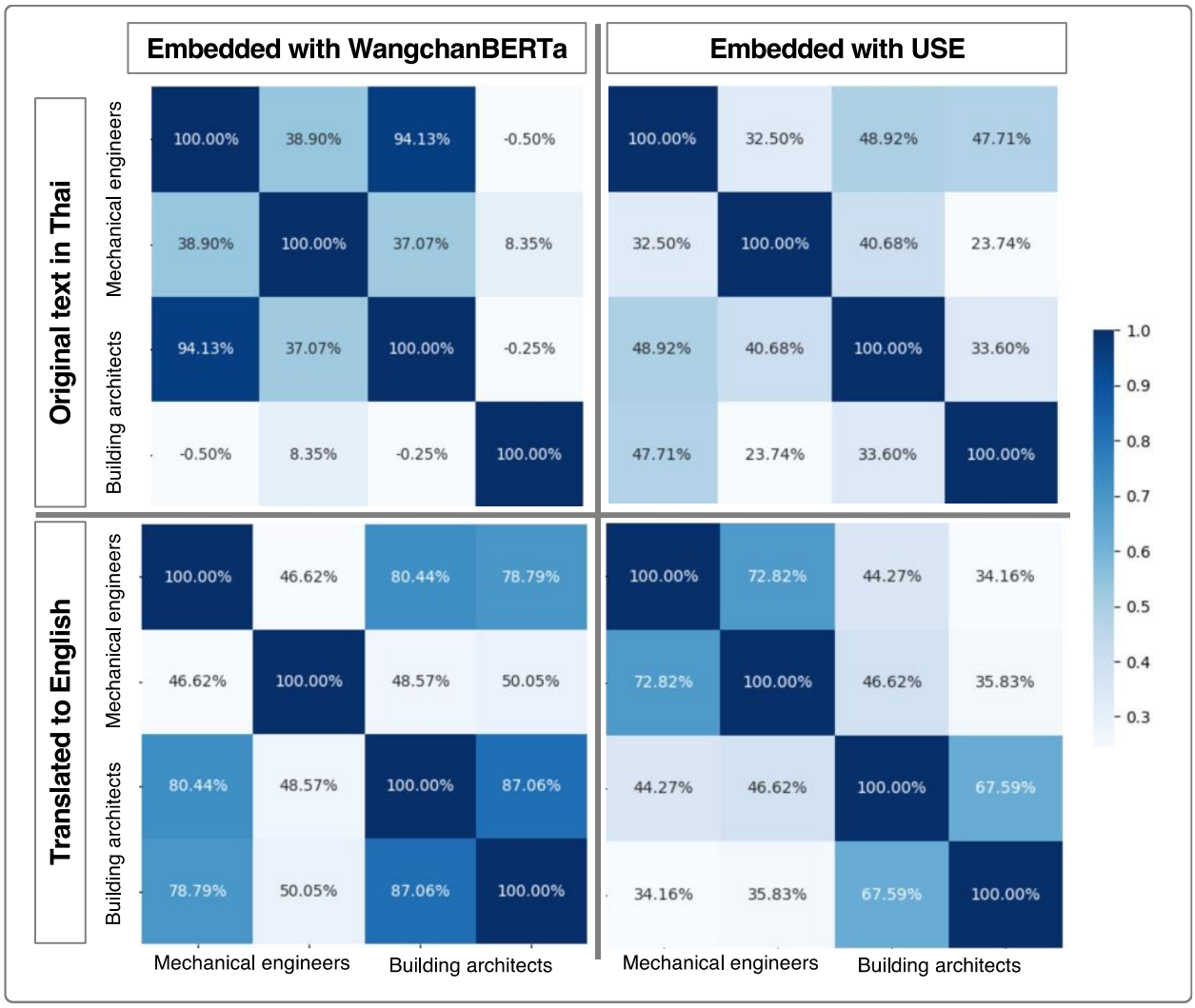

Figure 1: Cosine Similarity Matrices at Stage One

First, we note that WangchanBERTa, when applied to Thai text in the top-left quadrant, reveals a surprising challenge in distinguishing between disparate occupational categories. This is evidenced by the unexpectedly high similarity scores, such as 94.13% between Mechanical Engineer and Building Architect. This pattern persists with English translations in the bottom-left quadrant, where we observe high similarity scores between different categories, such as 80.44% and 78.79%, pointing to a consistent difficulty in separation despite the language transition.

Second, USE presents a stark contrast to the performance of WangchanBERTa - particularly when analysing English-translated text. While USE's capacity to differentiate between Mechanical Engineer and Building Architect in Thai is not without its faults (e.g., higher similarity scores of 48.92% and 40.68% on the top-right quadrant), it still outperforms WangchanBERTa' s attempt at the same task. The divergence becomes even more pronounced when the text is translated to English in the bottom-right, where USE's proficiency is highlighted by significantly lower inter-category similarity scores, e.g., 34.16% and 35.83%. Furthermore, USE aptly demonstrates its

capacity to recognize and reflect the semantic cohesion within the same job category, evidenced by robust intra-category similarity scores of 72.82% and 67.59%. These numbers not only demonstrate USE's more superior capability to separate distinct job categories but also shed light on the deficiencies of WangchanBERTa, which comparatively fails to offer the same level of clarity in stage one.

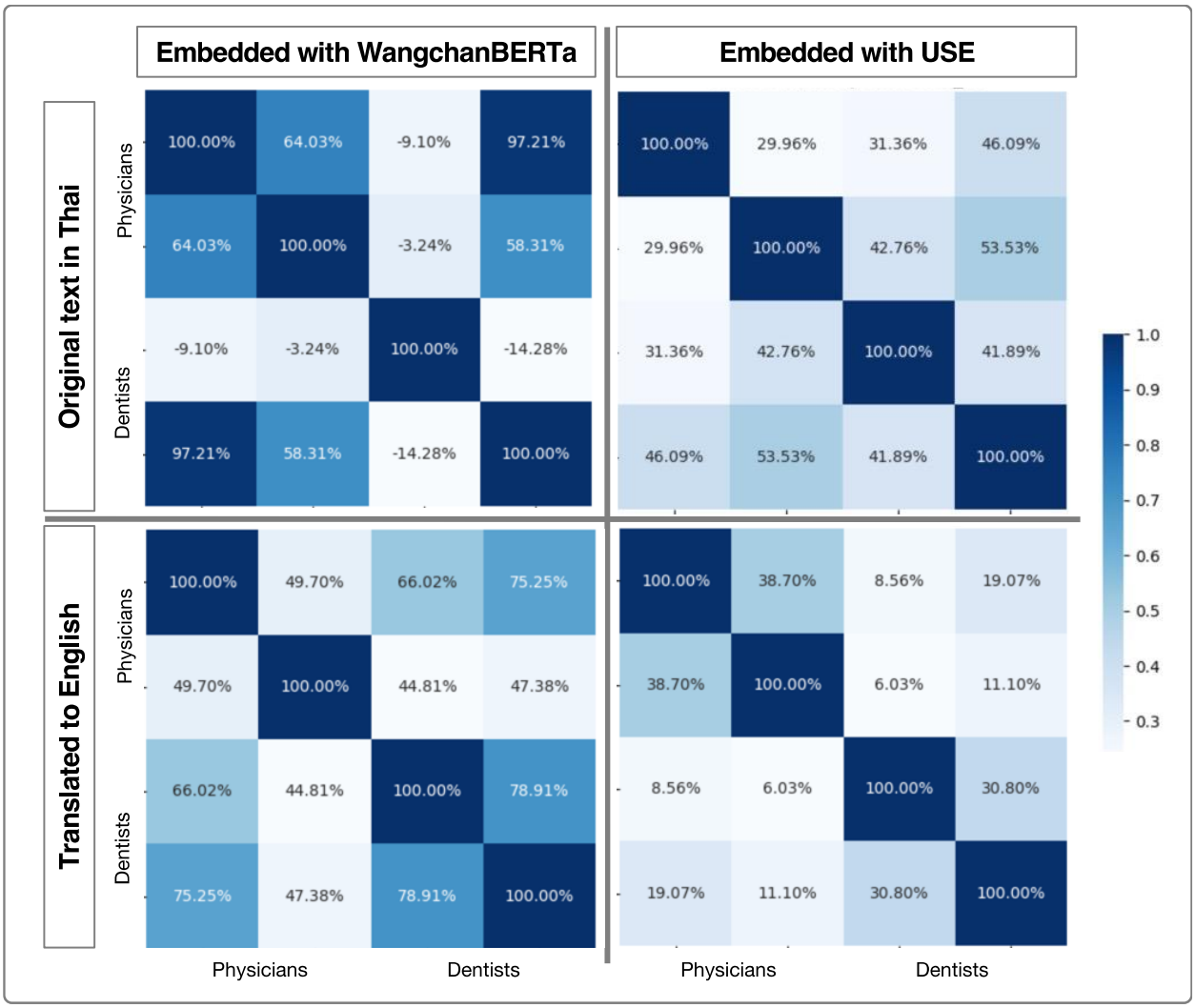

Figure 2: Cosine Similarity Matrices at Stage Two

The second stage of our analysis further interrogates the capacity of WangchanBERTa and USE to discern between job titles within the same professional field but with distinct specialties. The top-left quadrant of Figure 2 (WangchanBERTa with Thai text) illustrates an inherent limitation of the model. It shows high similarity scores, i.e., 97.21%, which indicates an inadequate capability to differentiate between the roles of 'Physician' and 'Dentist'. The English translation (the bottom-left) does little to mitigate this challenge, as evidenced by elevated similarity scores between the two professions, i.e., 75.25%.

In contrast, the performance of USE with the original Thai text (the top-right) offers an adequate differentiation, with lower similarity scores of 31.36% and 42.76% between 'physician' and 'dentist', suggesting a better grasp of the subtle differences between these medical roles. This understanding is further enhanced with the English translation (the bottom-right), where the inter-category similarity drops to 8.56% and 6.03%. This signals a robust ability of USE to distinguish between closely related job titles when processed in English. These results from stage two not only highlight the comparative strengths of USE, especially with English-translated text but also indicate the particular challenges faced by WangchanBERTa in distinguishing between similar yet distinct job categories.

The final stage of our embedding competition challenges the WangchanBERTa and USE with a sophisticated test: *distinguishing among three closely related accounting professions*. The heat maps for WangchanBERTa with Thai text (the left of Figure 3) show an inability to effectively differentiate between 'accountant', 'accounting associate professionals', and 'accounting and bookkeeping clerks'. The result exhibits high similarity scores across these categories, such as 93.08%, between 'accountant' and 'bookkeeping clerks'. The transition to English text (the right panel) does not significantly improve WangchanBERTa' s performance, with similarly high scores suggesting an overarching challenge in distinguishing closely related categories.

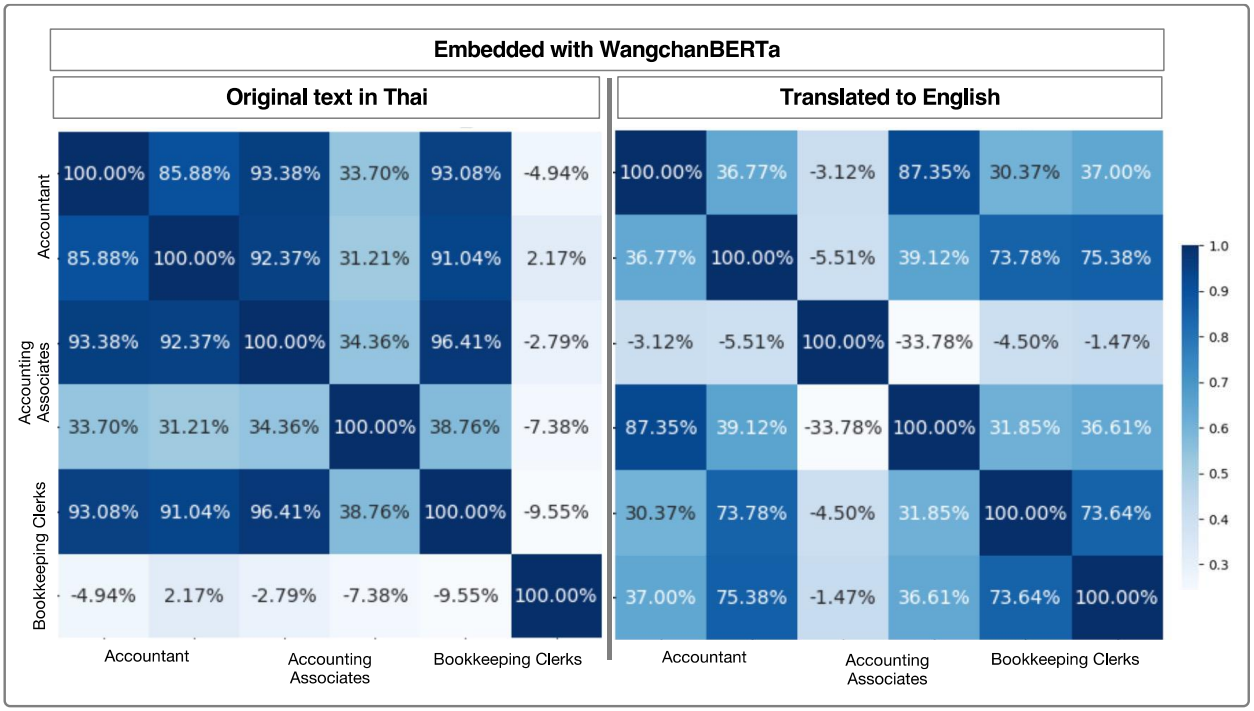

Figure 3: Cosine Similarity Matrices of WangchanBERTa at Stage Three

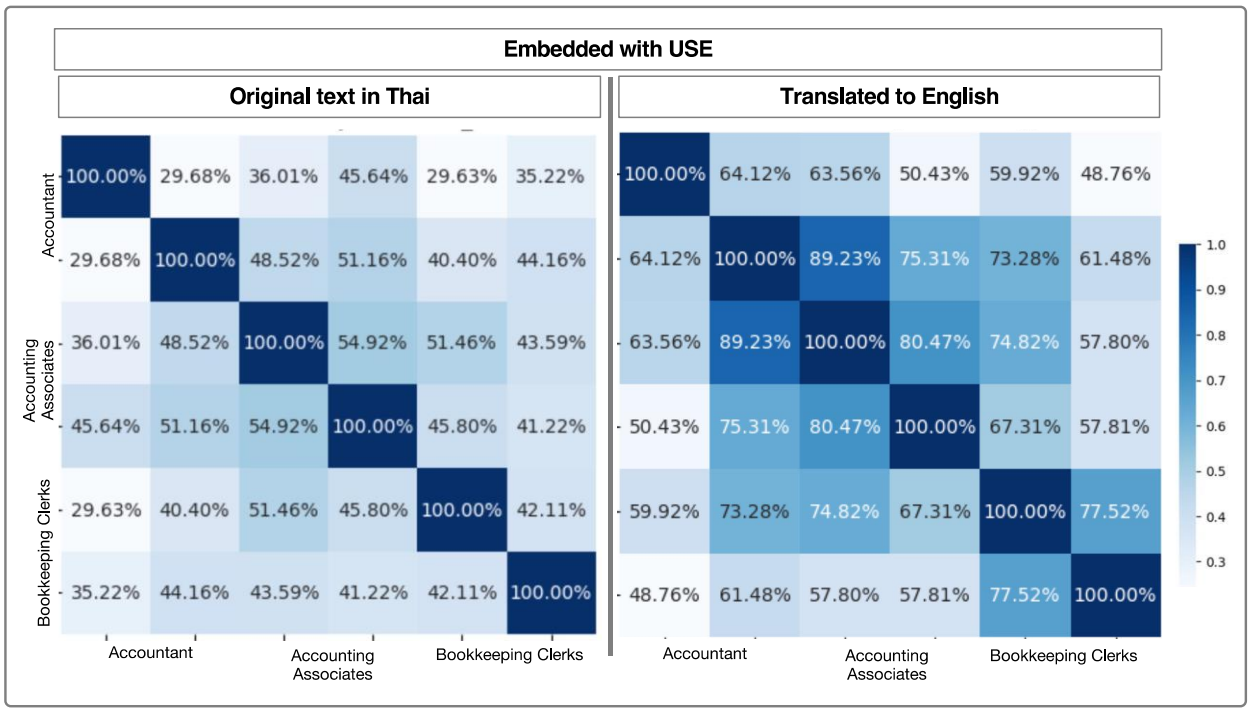

Figure 4: Cosine Similarity Matrices of USE at Stage Three

While the Universal Sentence Encoder' s (USE) performance with Thai text is not as exemplary, it manages to avoid the pitfall of high cosine similarity across all job titles irrespective of their ISCO categories—a significant issue noted in the WangchanBERTa' s performance with Thai text. This indicates that while not ideal, USE's performance with Thai text does not conflate distinct occupational roles into indistinguishable clusters, an essential baseline for effective job title classification. Moreover, when the analysis shifts to the English-translated text (the right of Figure 4), USE's performance gains notable traction. The model shows a commendable ability to distinguish between closely related professions, as evidenced by the substantial drop in similarity scores between different ISCO categories. For instance, the inter-category similarity between 'Accounting Associate Professionals' and 'Accounting and Bookkeeping Clerks' registers lower scores, while still maintaining high scores among their respective ISCO codes. This is indicative of a decent distinction and confirms its proficiency in English text embedding for nuanced classification tasks.

Given these observations, the comprehensive three-stage analysis has revealed that the Universal Sentence Encoder, particularly when processing English-translated text, consistently outperforms WangchanBERTa in distinguishing job titles across various levels of occupational linguistic challenge. The lower inter-category similarity scores achieved by USE in English highlight its superior capability to capture the fine-grained distinctions necessary for accurate job classification.

Therefore, to optimize our text classification methodology, we will adopt the Universal Sentence Encoder in conjunction with English-translated text to ensure the highest level of precision in our machine learning models. This decision is informed by the clear empirical

evidence of USE's robust performance across all testing stages, proving its suitability for our task of classifying web-scraped job posts into their respective ISCO codes.

3. THE (ACCURACY) RACE OF MACHINES

3.1. Methodology: Workflow and Metrics

In the expedition of finding the best machine learning models for job title classification, our methodology commences with the following steps:

i. Cleaning of text data, specifically job titles.
ii. Applying the 'majority rule' as our first decision rule. This is an approach that addresses the challenges inherent in our dataset explained in the next section, in order to resolve ambiguities and inconsistencies in the manual ISCO code assignment process. This step is crucial for ensuring data integrity and creating a reliable foundation for subsequent model training.
iii. Training a variety of models, including the Cosine Similarity Classifier, Hierarchical Cosine Similarity Classifier, Random Forest, eXtreme Gradient Boosting, Neural Networks, and BERTopic. These models are meticulously trained and then validated using the 5-fold validation technique to ensure robustness and reliability.

Accuracy is the chosen metric for evaluation, primarily due to the extensive range of categories (415 categories of 4-digit ISCO codes) involved in this study. Accuracy rate provides a straightforward, aggregate assessment of each model's ability to correctly classify a wide range of job titles into these diverse categories. On the contrary, utilizing metrics such as recall, precision, or F1 score would require a detailed and complex analysis for each individual ISCO code, which would not only be time-consuming but also impractical given the colossal number of categories.

3.2. Key Data Issue: Human Errors

Our primary dataset for model training, sourced from the Department of Employment (DoE) under the Ministry of Labour, Thailand, presents a unique set of opportunities as well as challenges. On the one hand, this dataset, meticulously compiled by the DoE and its regional branches, encompasses a comprehensive, administrative record of job postings across Thailand (including those postings directly with DoE or with selected job websites. It details company names, job titles, ISCO codes, and qualifications required. On the other hand, the issue arises in the method of data categorization. In fact, the 4-digit ISCO codes that we observe in the data were manually assigned to each job title by DoE personnels (from different regional offices). This introduces a degree of subjectivity and inconsistency in the classification process. We have the situation whereby the same occupation title is assigned different 4-digit ISCO codes. (Note also

that, one ISCO code can be valid for different occupational titles. But in this case, it does not pose an issue to the classification as we explain below.)

The challenges posed by this dataset could be a deal breaker for classification tasks. Ideally, the relationship between job titles and ISCO codes should be *many-to-one*, where a single 4-digit ISCO code encapsulates a range of similar job titles. Contrarily, our analysis revealed a *many-to-many* relationship, a significant anomaly that could severely undermine the efficacy of text classification efforts. This discrepancy likely stems from varying interpretations and definitions of ISCO codes across different regional offices and even among individual staff members within the same office. Such variations in judgment lead to a fragmented and inconsistent dataset, where the same job title might be associated with multiple ISCO codes.

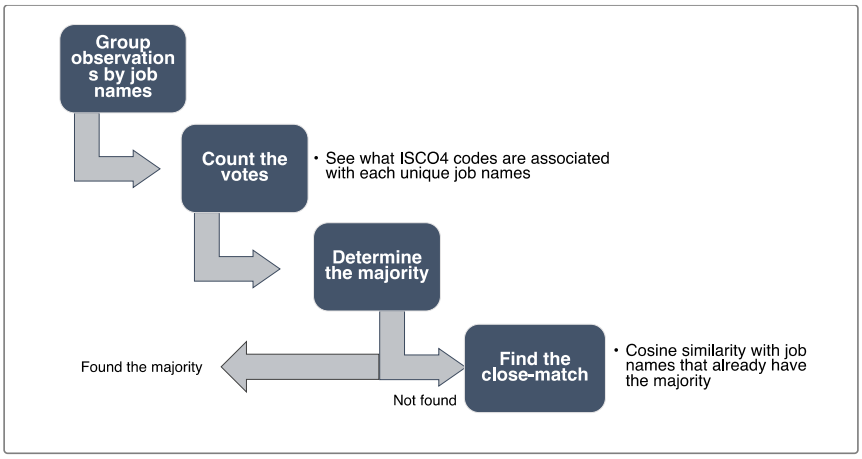

Figure 5: The Process of Determining the Majorities

(a) Applying majority rule: To confront the challenges inherent in our dataset, we devised a robust two-pronged solution strategy (Figure 5). The first phase involved instituting a 'majority rule' approach to sanitize the data. This process entailed a meticulous grouping and counting of 4-digit ISCO code variants corresponding to each unique job title. We then identified instances where a particular ISCO code dominated by a significant margin – a clear majority. These instances were deemed as 'cleaned data', serving as a benchmark for further analysis.

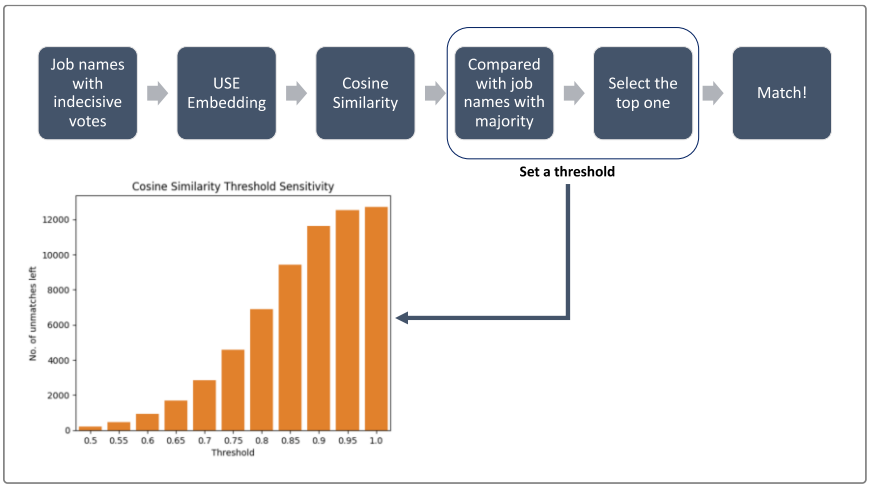

Figure 6: The Close-Match Approach

(b) Close-match approach: In the second phase, we turned our attention to the unique job titles that still lack a clear-cut, majority ISCO code. Here, we implemented the close-match approach (Figure 6). Each title was embedded using the Universal Sentence Encoder (USE), an aforementioned tool adept at capturing the nuanced semantic essence of textual data. We then computed the cosine similarity of these embedded unclean titles with the benchmark data, also transformed via USE. This comparison aimed to identify the closest match for each unclean job title within our benchmark group, based on the highest cosine similarity score that surpassed a calibrated threshold.

This threshold, carefully optimized, served as a critical determinant in ensuring both the accuracy of our matches and the meaningful expansion of our dataset. Job titles from the unclean group that found a counterpart in the benchmark, with cosine similarity scores exceeding the threshold, were assigned the cleansed ISCO code of their corresponding match. This strategic approach allowed us to extend the clarity and precision of the *majority rule* to the remaining ambiguous titles, further refining our dataset. By employing this method, we ensured that each job title, irrespective of its initial classification ambiguity, was accurately aligned with one 4-digit ISCO code, which is the most appropriate ISCO code. By doing so, we enhanced the overall integrity and applicability of the data for labour market analysis, while limiting the loss in the volume of original data.

3.3. Text Classifications

3.3.1. Cosine Similarity Classifier

In our pursuit of effective text classification, particularly in the context of matching web-scraped job titles to ISCO codes, the Cosine Similarity classifier emerges as a foundational tool.

We designed a text classifier that operates on the principle of utilizing the semantic meanings that were embedded in a high-dimensional space and the geometric calculation of Cosine, a straightforward concept yet well-suited to text data that has been transformed into vector representations. By embedding both the job titles and ISCO descriptions using our chosen Universal Sentence Encoder, we convert these textual entities into dense vectors that capture the essence of the original text remarkably well.

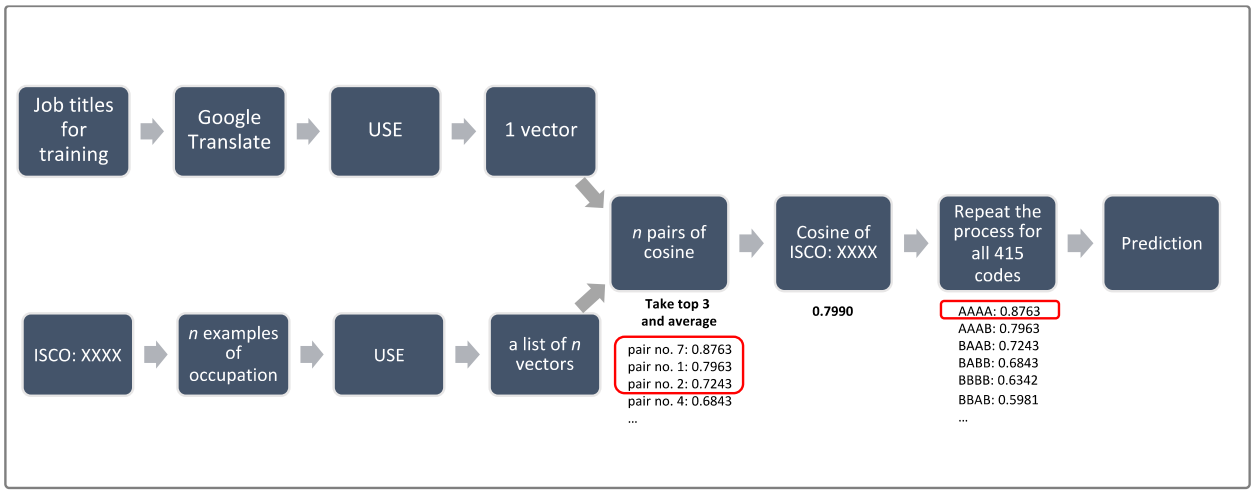

Figure 7: Cosine Similarity Classifier Workflow

As shown in Figure 7, the process of classification involves calculating the cosine similarity between the vector representation of a given job title input and each ISCO description vector. Cosine similarity, whose mathematical expressions are denoted and discussed in an earlier section, is selected to be the core engine of this design. A key rationale lies in its capability to effectively quantify how closely related two text documents are in terms of their content and context. In our application, each comparison yields a cosine similarity score, that ranges from -1 to 1, which is then used to rank all the ISCO descriptions paired with the job title.

The classification decision is based on this ranking: the ISCO description with the highest cosine similarity score to the job title is deemed the closest match and, therefore, the predicted ISCO for that job title. This approach harnesses the semantic richness of USE embeddings, ensuring that the classification is not just based on superficial text matching but on a deeper understanding of the contextual and semantic relationships between job titles and occupational descriptions.

3.3.2. Hierarchical Cosine Similarity Classifier (Nested Models)

Building upon the foundation of the Cosine Similarity Classifier, our Hierarchical Cosine Similarity Classifier introduces a more nuanced approach, particularly beneficial for classification scenarios where categories are hierarchical, such as our 415 categories of ISCO codes. This classifier employs a nested model structure, where the classification process is performed in

hierarchical levels, each focusing on a specific level of ISCO digits. Note that ISCO codes contain up to 4 digits, with the first digit is the most aggregated level.

At the heart of this classifier is a hierarchical process that sequentially matches job titles to ISCO codes, starting from the broadest category (the first digit) and progressively narrowing down to the specific occupation (the fourth digit). The first digit of ISCO codes represents major occupational groups. In the initial stage, we aggregated ISCO descriptions, creating a set of descriptions for each major group (first digit ISCO) that encompass all sub-category descriptions. For each first-digit category, the classifier computes the cosine similarity between the job title vector and each individual ISCO description vector within that category. The category with the description yielding the highest cosine similarity score is then used to represent that particular first digit.

Once the classifier identifies the most likely first digit based on the highest cosine similarity score in that category, it proceeds to the next level, focusing only on the ISCO codes that share this predicted first digit. This refinement significantly streamlines the classification process, as the classifier now only considers the relevant subset of ISCO codes for further matching. The classifier then repeats this process for the second, third, and fourth digits, each time narrowing the focus and recalculating cosine similarities within the increasingly specific subsets of ISCO descriptions.

This hierarchical, step-by-step approach ensures a more targeted and efficient classification process. By computing cosine similarities individually at each digit level and progressively filtering the ISCO codes, this classifier not only enhances accuracy but also optimizes computation time.

3.3.3. Random Forest

Apart from the two manually designed classifiers/workflows introduced in previous sections, a number of supervised machine learning algorithms are also thoroughly explored in this paper to heighten the possibility of maximizing prediction accuracy. The first candidacy is Random Forest, one of the highly well-known machine learning algorithms used in a wide array of classification tasks due to its robustness and accuracy, as well as its ability to capture complex, non-linear relationships between features and classes. It operates by constructing multiple decision trees during training. Each decision tree in the forest, to its best knowledge, makes a prediction based on a random subset of these textual features, and the final classification is determined by a majority vote among all trees. This ensemble method is particularly effective in handling large datasets with high dimensionality, making it suitable for text classification tasks where features are often colossal and complex.[15] Moreover, its bagging nature reduces the risk of overfitting, a common challenge in text classification, and improves generalization to new and unseen data

[15] The adaptability and efficacy of the Random Forest algorithm are illustrated through diverse applications across various text classification studies. Ramayanti and Salamah (2018) focused on classifying Twitter data related to Indonesia's Ministry of Marine Affairs and Fisheries, aiming to categorize tweets as complaints or non-complaints. Their approach demonstrated Random Forest's capability in accurately analysing and classifying, even with nuanced social media content, achieving a notable classification accuracy, 95.6% to be precise. Bouaziz et al. (2014) tackled

3.3.4. eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is a highly sophisticated machine learning algorithm, particularly effective in the field of classifications. Building upon the concept of ensemble learning, similar to Random Forest, XGBoost takes a more refined approach by employing gradient boosting frameworks. Unlike Random Forest, where trees are built in parallel, XGBoost constructs them sequentially, with each new tree correcting the errors made by the previous ones, making each tree progressively better. This method involves optimizing a loss function, where the algorithm focuses on the gradients (errors) and continuously works to minimize these through the addition of new trees.

One of several qualities that makes XGBoost a qualified candidacy for our study is its ability to handle a wide range of data sparsity, a challenge commonly found in text data, with a robust mechanism to avoid overfitting. This is achieved through regularized learning, which is applied to both the tree structure and the leaf weights. XGBoost also efficiently manages computing resources, making it faster and more scalable than many other algorithms. Moreover, its ability to handle missing data and its flexibility to customize the optimization objectives and evaluation criteria make it particularly suitable for complex text classification tasks like ours.[16]

3.3.5. Deep Learning

Deep learning, especially when integrated with advanced embedding like USE, offers a stellar approach to text classifications, especially in our scenario where we have over 400 ISCO codes to distinguish. The key to deep learning's superiority lies in its ability to learn hierarchical

---

the challenge of short text classification, a task complicated by sparseness and lack of context. They combined data enrichment with semantic analysis within the Random Forest framework, significantly enhancing classification accuracy by 34% over traditional methods. This advancement underscored the potential of integrating semantic understanding into machine learning models for more nuanced text analysis. Meanwhile, Sun et al. (2020) sought to improve the traditional Random Forest algorithm for text classification by introducing a weighted voting mechanism and employing the BERT word vector model for text representation. This approach aimed to enhance the quality and diversity of text features, addressing the limitations of traditional feature extraction methods. Their results showed superior performance in text classification, highlighting the benefits of advanced feature extraction techniques when combined with calibrations in Random Forest.

[16] Numerous studies exhibit the superiority of the XGBoost application in challenging text classification tasks. Ghosal and Jain (2023) developed a framework for distinguishing between depression and suicidal risk content on a Reddit dataset, utilizing fastText embedding for contextual analysis and TF-IDF vector for term relevance. Their approach achieved an AUC of 0.78 and a weighted F1-score of 0.71, showcasing the efficacy of combining fastText embedding with XGBoost in analysing complex emotional content. In a different application, Hendrawan, Utami, and Hartanto (2022) compared the performance of Word2vec and Doc2vec embedding methods on unbalanced review text data, using XGBoost for classification. Their study revealed that both Word2vec and Doc2vec, when paired with XGBoost, could effectively classify unbalanced datasets, achieving an average F1 Score of 0.9342 and 0.9344, respectively. This comparison highlighted the suitability of embedding and XGBoost cooperation for processing unbalanced data. Empirical evidence also illustrates the capability of XGBoost in different terrains. Tiwari and Agrawal (2022) conducted a comparative analysis of machine learning methods for hate speech recognition on Twitter. Using the Davidson dataset, they found that XGBoost, combined with TF-IDF transformer embedding, outperformed other models, achieving an impressive accuracy of 94.4%. This study underscored the effectiveness of TF-IDF embedding with XGBoost in accurately detecting hate speech rooted in short, nuanced social media text data.

representations of data. Unlike ensemble methods like Random Forest and XGBoost, which rely on decision trees, feed-forward neural networks use layers of neurons where each layer's output is the input for the next. This structure allows the network to learn complex patterns in data, making it particularly effective for high-dimensional tasks like text classification. When combined with USE, the advantage is expanded twofold: firstly, it leverages the deep semantic understanding from USE and secondly, the high-dimensional vectors are utilized by the hierarchical learning capability of neural networks, leading to more accurate and versatile text classification as several academic studies have demonstrated.[17]

In our design, as depicted in Figure 8 in the appendix, the architecture includes progressively narrowing the number of nodes in the dense layers, moving from the initial broader layer to more focused subsequent layers. This tapering serves a strategic purpose. By starting with a wider field, the network can capture a vast array of features and patterns in the data, as intended with the embedding mechanism of USE. As the model progresses through the layers, it begins to refine these features, prioritizing the most relevant information for classification. In addition, with the integration of techniques like Batch Normalization and Dropout, the model enhances its ability to generalize, reducing the likelihood of overfitting despite the high dimensionality of the input data. The final output layer employs a softmax activation function to yield a probability distribution across all 415 ISCO categories.

3.3.6. BERTopic (Topic Modelling)

Topic modelling is a prevalent technique in the field of NLP and has numerous applications in text classification, offering a way to discover latent themes from large collections of documents. The techniques to perform topic modelling vary in a broad spectrum, from more conventional methods like Latent Dirichlet Allocation (LDA) to far more cutting-edge algorithms like BERTopic, which transcends the limitations of its predecessors like LDA by leveraging contextual embeddings from models like BERT or USE. These embeddings capture deep semantic relationships within the text, enabling a more nuanced and contextually rich discovery of topics.

Furthermore, BERTopic allows users to incorporate labels during the topic modelling process. This feature signifies its unique semi-supervised capability, allowing integration of domain-specific knowledge, while preserving the liberation nature of unsupervised machine learning: being able to learn patterns and relationships between complex textual vectors and classes without hyper-fixations on strict, annotated inputs. By doing so, the model's relevance and accuracy are enhanced, especially in scenarios with a vast number of categories like ours and so

[17] Effectiveness of this combination has been utilized in various research papers, including an outstanding study done by Khandve et al. (2022). They explored hierarchical transfer learning approaches for long document classification using pre-trained USE and BERT. Their methodology involved slicing input data into chunks and passing them through BERT and USE. The embedded chucks will then be conveyed to shallow neural networks like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN). The same design was tested on six benchmark datasets. Their approach demonstrated that USE combined with CNN/LSTM outperforms standalone models.

many other real-world applications that would not be able to materialize without the help of cutting-edge techniques like BERTopic.[18]

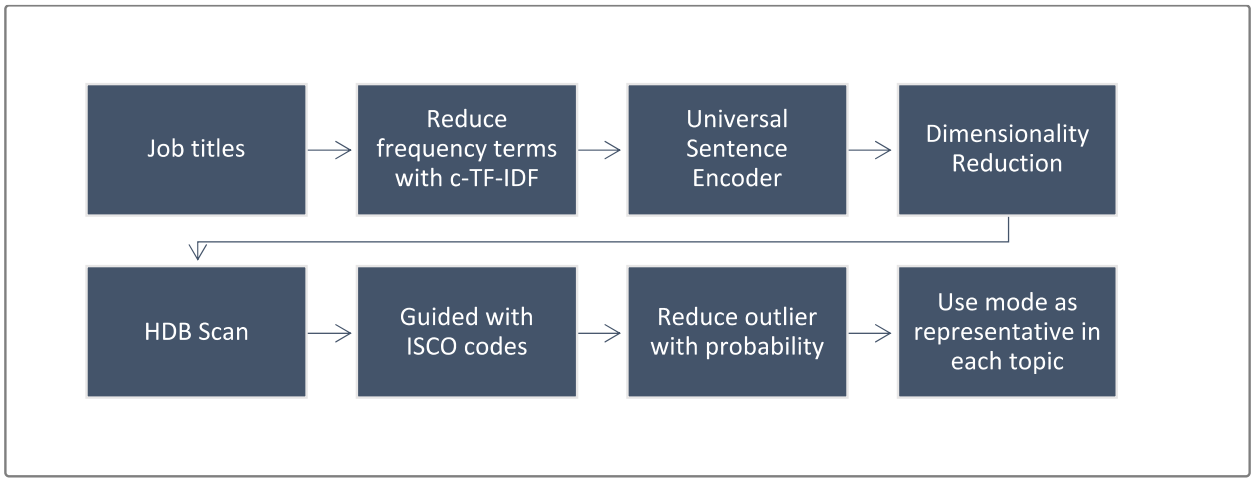

Figure 9: The BERTopic Workflow

The BERTopic workflow, as shown in Figure 9, commences with the raw job titles that are initially refined through a frequency-based c-TF-IDF process to emphasize the more distinctive terms. Subsequently, these refined titles are encoded using the Universal Sentence Encoder, identical to other classifiers. The encoded data undergoes dimensionality reduction to further streamline the modeling process. HDBSCAN then clusters the data, with the assistance of ISCO codes as guiding labels, shaping the model to the right clustered gravitations. Probabilities are calculated to minimize the impact of outliers, ensuring that the most representative mode of each topic is used to define the cluster. This semi-supervised approach balances the need for domain-specific insights with the broad pattern recognition capabilities of unsupervised learning, cultivating the sophisticated classification that aligns closely with the complex nature of job titles and their corresponding ISCO codes.

[18] Rey-Moreno et. al. (2023) explored unstructured, natural-language data from Airbnb and hotel stays, comprising 12,236 Airbnb sentences and 12,200 hotel sentences collected from 2018 to September 2021. BERTopic was used to identify latent themes in the narratives. The insights derived helped provide anecdotal evidence of varying priorities among Airbnb and hotel guests. For example, Airbnb guests frequently mentioned the uniqueness of the accommodation, the personal touch provided by hosts, and the authenticity of the local experience. These aspects were frequently mentioned in positive reviews, highlighting their importance in driving guest satisfaction in the Airbnb context. In contrast, for hotel stays, BERTopic identified different priorities among guests. Key themes included the professionalism of hotel staff, the quality and professionalism of hotel staff, the quality and comfort of rooms, and the efficiency of services like check-in and check-out processes. These factors were often cited in reviews that expressed high levels of trust in the hotel services. This information, which was unveiled by BERTopic, is invaluable for service providers in the hospitality industry as it could help tailor services to meet specific needs and develop targeted marketing strategies.

4. RESULT DISCUSSION

4.1. Accuracy Comparison

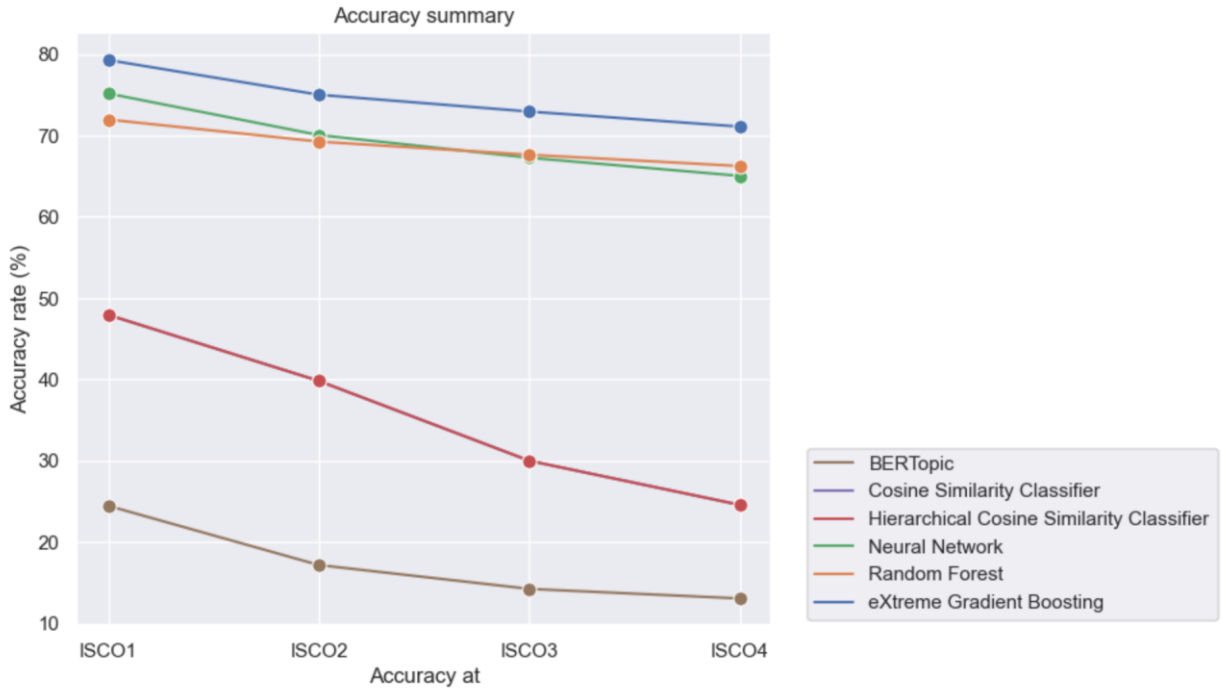

Figure 10: Accuracy Summary of All Models from the 1st to 4th digits

Figure 10 illustrates the accuracy rates of each model, from the first digit of ISCO to its fourth digit, measured by the 5-fold validation technique. The eXtreme Gradient Boosting emerged as the clear frontrunner, demonstrating exceptional performance by leveraging its ability to capture complex and non-linear patterns among the diverse lexicons of the labor market context with semantic insights provided by USE embeddings and diverse ISCO categories to achieve the most impressive results. Its performance, particularly at the more granular fourth digit level, is exemplary, with a 71% accuracy rate. The high degree of precision, especially in the context of detailed classification, can be attributed to the inherent advantages of its ensemble and optimizing nature. This aggregating method is particularly effective in managing the non-linear complexities of our high-dimensional, vectorized text data to their corresponding ISCO codes. We observe it by minimal losses of accuracy between broader digits and more precise digits. Moreover, its ensemble nature helps to mitigate overfitting. By leveraging multiple decision trees, this model ensures a more generalized and robust prediction, avoiding the trap of being overly tailored to the training data. This aspect is crucial in our context, where the ability to generalize across a vast array of job titles and corresponding ISCO codes is crucial. The consistent performance of XGBoost across all levels of classification demonstrates its suitability for our text classification task, where the dataset is large and varied in classes.

The Random Forest model, while also utilizing an ensemble approach, showed commendable robustness with accuracy rates closely trailing those of XGBoost, achieving 72% at the first digit and 66% by the fourth digit. Although it shares a similar ensemble nature with XGBoost, it lacks the additional optimization capabilities that XGBoost employs, which could be the reason for the slight discrepancies in performance between the two. The Random Forest model still proves to be highly effective, especially considering its ability to generalize well across the high-dimensional, vectorized text data linked to ISCO codes.

Meanwhile, the Neural Network model, designed with a streamlined arrangement of dense layers and regularization techniques such as Batch Normalization and Dropout, presents an interesting case study in the dynamics of accuracy concerning the granularity of ISCO codes. The initial promising accuracy of 76% at the first-digit level underscores the model's capability to capture broad occupational categories. However, as we delve into the more refined classification towards the fourth digit, the accuracy rates decline to 65%. This somewhat steep decrease is indicative of the model's limitations in deciphering and maintaining the capability to distinguish the job titles as the classification becomes more specific. The decline in accuracy may reflect the need for subtle refinements in the model's existing framework rather than an overhaul. This could involve minor architectural adjustments or incremental enhancements to the training process, which could be explored in future iterations.

The Cosine Similarity Classifier and the Hierarchical Cosine Similarity Classifier, both designed to leverage similarities in text data, showed comparable and modest performance across the ISCO digit levels. Both models began with an accuracy of approximately 48% at the first digit and decreased to around 25% by the fourth digit. This decline implies that the classifier, while proficient in capturing overarching themes within job titles, may not be as effective in distinguishing the finer details that differentiate one job from another at a more granular level. The parallel trends in their performance also suggest that building a pipeline of prediction to imitate the hierarchical structure of the ISCO categories does not yield a marginal advantage. The design of both cosine similarity approaches, while straightforward, appear to lack the depth needed to fully capture the complex relationships and differentiations among highly specialized and vastly varieties of real-world job titles.

The BERTopic approach stands distinct as the sole semi-supervised model in our study. Its performance demonstrates the challenges inherent in applying comparatively unsupervised learning to such a structured and detailed classification task. With accuracy rates starting at 24% at the first digit and declining to 13% at the fourth digit, BERTopic's performance is considerably lower than the supervised models. This drop in accuracy underscores its limitations in accurately mapping the complex job title descriptions to their appropriate ISCO codes without direct supervision. The reliance on thematic similarity within topics, while innovative, may not be sufficient to capture the subtleties required for precise ISCO classification, particularly when dealing with the vast and diverse array of job titles present in the data set.

In summary, our extensive exploration across various models for ISCO code classification has highlighted the superior capabilities of the eXtreme Gradient Boosting (XGBoost) due to its

advanced ensemble techniques and optimization algorithms. The Random Forest remains a strong contender, showcasing robust performance across all classification levels. Meanwhile, both the Cosine Similarity Classifier and the Hierarchical Cosine Similarity Classifier, despite their theoretical appeal, fall short in practical application when the classification demands greater specificity and depth. The BERTopic model, while offering a unique unsupervised approach, illustrates the challenges of applying such techniques to highly structured classification tasks. These outcomes underline the challenges of bridging the gap between the idealized ISCO guidebook categories and the complexity of real-world job titles.

Given these findings, the decision to proceed with the eXtreme Gradient Boosting (XGBoost) model for our classification task would be a strategic one, informed by its outstanding performance and practical advantages. The adaptability and resilience of the XGBoost model, especially in conjunction with the semantic depth provided by the USE, position it as the most fitting choice for our project's objectives. It promises not only accuracy in classification but also the capability to manage the volume and variety inherent in our dataset. Moreover, the logistical aspects of deploying XGBoost, considering its manageable model size and resource efficiency, align well with our operational constraints and project timeline.

4.2. LLM Integration

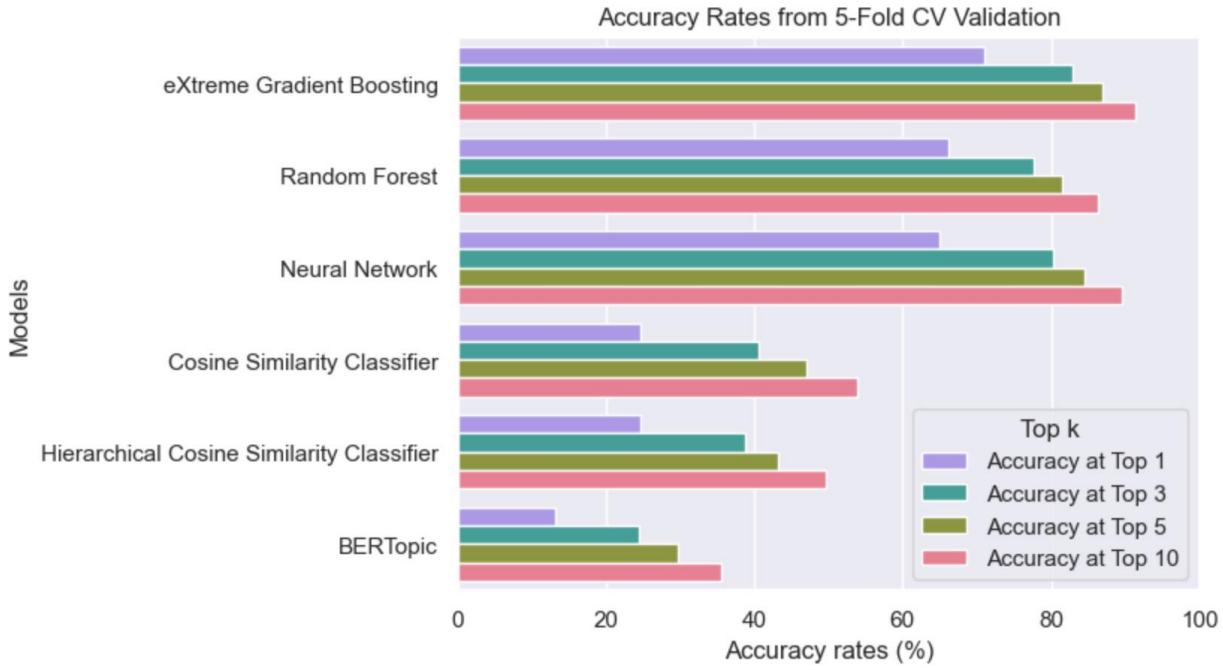

Figure 11: Accuracy Summary of All Models at top 1, 3, 5, and 10 predictions

In the pursuit of enhancing the accuracy of our eXtreme Gradient Boosting (XGBoost) model, we noted a significant discrepancy in performance between the top 1 accuracy (71%) and

the top 3 accuracy (83%) as shown in figure 11. This 12% increase in the top 3 accuracy indicates that while the XGBoost model often ranks the correct ISCO code within its top three predictions, it does not consistently identify it as the most probable. This observation suggests a near-miss in classification that could potentially be rectified. To address this, we explored the integration of Phi-3, one of the state-of-the-art Large Language Models (LLMs) developed by Microsoft, renowned for its compact size yet robust performance in complex language processing tasks.

The integration process involved employing Phi-3 at the end of the pipeline to re-assess the top three predictions from XGBoost for each job title. Phi-3 was tasked with analyzing the context and details within the three ISCO codes proposed by XGBoost accompanied by their respective definitions from the ISCO guidebook, to determine which code best matches the job titles. By leveraging its advanced natural language understanding, Phi-3 was expected to enhance decision accuracy by pinpointing the most appropriate ISCO code from the narrowed-down potential options, thereby reducing the instances of near-misses and enhancing the overall classification precision.

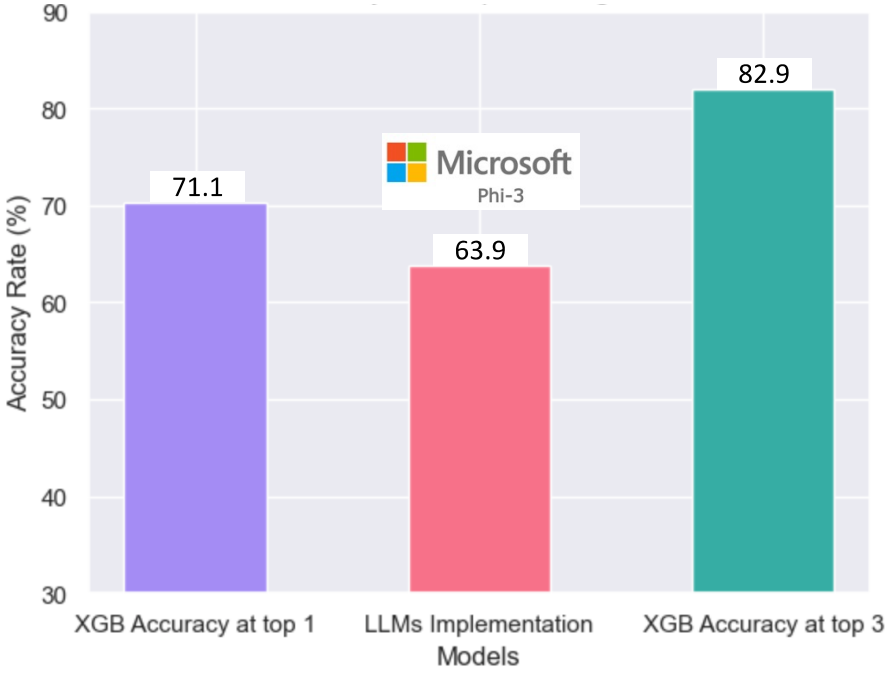

Figure 12: Accuracy rate from Phi-3 LLM integration with XGBoosting Classifier

Despite the theoretical capabilities of Phi-3, the practical integration yielded suboptimal results as illustrated in figure 12. Instead of improving the XGBoost model's predictions accuracy closer to the top 3 prediction accuracy, the integration saw a decline to 64% from the standalone 71% accuracy of the pure XGBoost model. This reduction was largely due to the LLM's re-evaluation of already correct predictions, where it introduced errors instead of rectifications. The

Phi-3 model sometimes failed to correctly interpret the context and nuances of the job descriptions against the ISCO codes, leading to less accurate outcomes.

This experience with Phi-3, although not as successful as anticipated, provides valuable insights into the complexities of integrating advanced LLMs into specific classification frameworks. However, given the decrease in performance with the integration of Phi-3, and considering the operational and computational efficiency, we will continue to utilize the XGB model without LLM enhancement for our classification needs. This decision is underscored by the need to maintain high accuracy and reliability in our ISCO code classification tasks, where the straightforward application of XGBoost already meets our requirements efficiently

4.3. Confusion Matrix

The aggregated confusion matrix for the eXtreme Gradient Boosting (XGB) model is depicted in Figure 11. The rows represent the actual ISCO categories at the first digit, while the columns depict the predicted ones. The values within the matrix show the proportion of predictions, with the diagonal cells indicating correct predictions and others showing which ISCO codes the model incorrectly predicted. A darker shade in these cells reflects a higher accuracy. This matrix is particularly insightful, as it sheds light on the model's strengths and, crucially, on potential systematic errors that could arise in the classification process. These figures are obtained from a 5-fold cross-validation to ensure a thorough assessment of its predictive power.

The superiority of the XGB model is particularly evident in the darker shades along the diagonal of the matrix, illustrating concentrated accuracy in correct predictions. Notably, the model excels with the Service and Sales Workers group (ISCO broad group 5), the Plant and Machine Operators and Assemblers group (ISCO broad group 8), and the Professional group (ISCO broad group 2). These groups are significantly represented in the dataset we intend to apply the model to, indicating that the model is well-calibrated to the types of job titles that it will encounter in practical use.

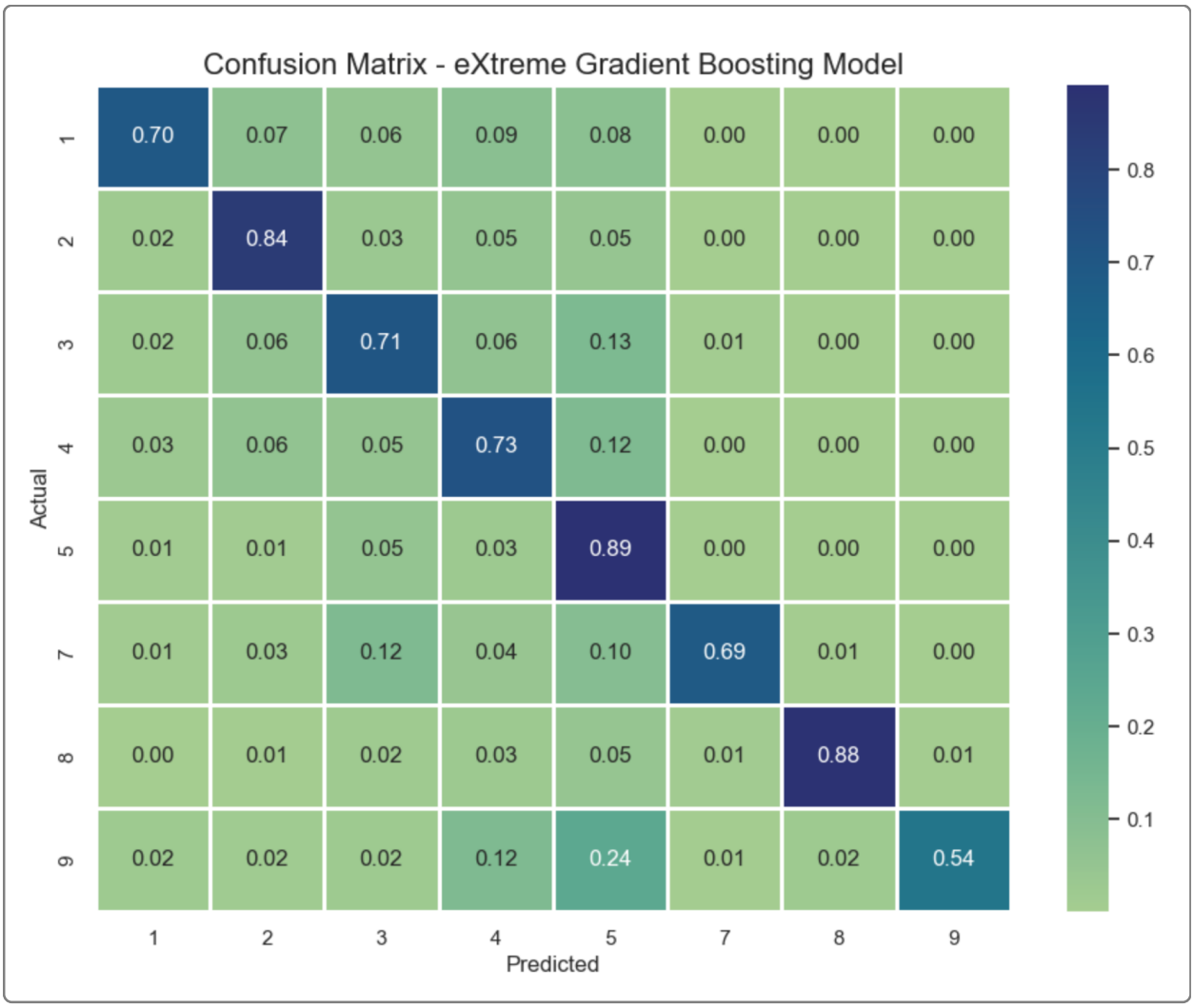

Figure 13: Confusion Matrix of the XGB model (illustrated at the first digit (ISCO 1) accuracy)

However, it is important to note that some ISCO categories demonstrate lower accuracy. For instance, there is a significant rate of misclassification where jobs in Elementary Occupations (ISCO group 9) are predicted as Service and Sales Workers (ISCO group 5), with an error rate of 24%. A notable addition to the error rate is accounted by the confusion between Domestic Helpers and Cleaners (classified under ISCO group 5, as ISCO 5151) and Elementary Occupations (ISCO group 9), where roles traditionally defined under specific classifications of domestic and cleaning services are inaccurately grouped under the broader and less specialized category of elementary workers. This error may stem from the human errors in ISCO classification of the job descriptions used in training the model, where tasks such as cleaning, assisting, and other low-skill duties are common to both categories but should be distinctly classified according to the ISCO guidelines.

Furthermore, the roles of Accounting Associate Professionals (ISCO 3313) from group 3 can be confused with those of Accounting and Bookkeeping Clerks (ISCO 4311) due to their involvement in financial documentation and accounting tasks in spite of their differential hierarchy in a job ladder. Similarly, the customer service-oriented roles of Shop Sales Assistants (ISCO

5223) could be inaccurately classified as Business Services Agents (ISCO 3339) for the fact that both jobs engage extensively in client interaction but in a different context. This misclassification may occur due to the model's struggle in differentiating the specific subtexts of the work environment and the scope of responsibilities that distinguish customer-facing roles in commercial sales from those in business service settings.

The accuracy for the Managers category (ISCO group 1), while not exemplary, may be influenced by the inherent ambiguity present in job listings and descriptions. Often, the specific nature of managerial roles, particularly in a job ad, is not fully specified, with job titles merely labelled as 'Manager' without further description. This lack of specificity poses a significant challenge for the XGB model, which relies on detailed data to distinguish between nuanced managerial functions. On the contrary, the job titles contain overwhelming information to the point where the jargons that indicate managerial levels are overshadowed. This can lead the model to latch onto contextual clues within the job title rather than the managerial terminologies that would indicate a higher level of responsibility, resulting in misclassification leaning towards major group 2 to 5.

4.4. Comparison of Granular Predictions across Models.

To get a better understanding of the supremacy of the XGB model in the context of the Thai labour market classification, rather than analysing the aggregated metrics, the 10 most frequent job titles from an online job post website are chosen to demonstrate the models' capability. As shown in Table 1, the translated job titles are listed in the furthest left column, where the predictions of each aforementioned model are laid out as follows:

Table 1: Predictions from all models

Translated Job Titles	Cosine Similarity Classifier	Hierarchical Cosine Similarity Classifier	Random Forest	eXtreme Gradient Boosting	Neural Networks	BERTopic
Salesperson	Glaziers (7125)	Biologists, botanists, zoologists and related professionals (2131)	Shop sales assistants (5223)	Shop sales assistants (5223)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Accounting officer	Valuers and loss assessors (3315)	Farming, forestry and fisheries advisers (2132)	Accountants (2411)	Accountants (2411)	Contact centre salespersons (5244)	Contact centre salespersons (5244)

Table 1: Predictions from all models

Translated Job Titles	Cosine Similarity Classifier	Hierarchical Cosine Similarity Classifier	Random Forest	eXtreme Gradient Boosting	Neural Networks	BERTopic
Customer service staff	Musicians, singers and composers (2652)	Musicians, singers and composers (2652)	Contact centre salespersons (5244)	Contact centre salespersons (5244)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Manager	Authors and related writers (2641)	Authors and related writers (2641)	Credit and loans officers (3312)	Credit and loans officers (3312)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Housekeeper	Hand and pedal vehicle drivers (9331)	Sociologists, anthropologists and related professionals (2632)	Domestic housekeepers (5152)	Domestic housekeepers (5152)	General office clerks (4110)	Contact centre salespersons (5244)
Procurer	Specialist medical practitioners (2212)	Specialist medical practitioners (2212)	Procurers (3323)	Procurers (3323)	General office clerks (4110)	Contact centre salespersons (5244)
Driver	Specialist medical practitioners (2212)	Specialist medical practitioners (2212)	Car, taxi and van drivers (8322)	Car, taxi and van drivers (8322)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Administrative officer	Authors and related writers (2641)	Authors and related writers (2641)	General office clerks (4110)	General office clerks (4110)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Technician	Authors and related writers (2641)	Authors and related writers (2641)	Process control technicians not elsewhere classified (3139)	Process control technicians not elsewhere classified (3139)	Contact centre salespersons (5244)	Contact centre salespersons (5244)
Foreman	Chefs (3434)	Chefs (3434)	Construction supervisors (3123)	Construction supervisors (3123)	Construction supervisors (3123)	Business services and administration managers not

Table 1: Predictions from all models

Translated Job Titles	Cosine Similarity Classifier	Hierarchical Cosine Similarity Classifier	Random Forest	eXtreme Gradient Boosting	Neural Networks	BERTopic
						elsewhere classified (1219)

The performance of the Cosine Similarity Classifier and the Hierarchical Cosine Similarity Classifier, while innovative in their conceptual design, appears to fall short in practice. Their performance demonstrated a tendency to correlate job titles with ISCO codes that depart significantly from their expected classifications. For example, both models assigned the role of 'Salesperson' to 'Glaziers (7125)' and 'Biologists, botanists, zoologists and related professionals (2131)', respectively. These dissociations may be indicative of their rigid design, grounded in the fixed occupational examples of the ISCO guidebook, which proved inflexible when confronted with the real-world data's broader set of jargon and terminologies. The Neural Networks and BERTopic classifiers, while sophisticated in their own right, recurrently categorized a broad spectrum of job titles under the singular ISCO code for 'Contact centre salespersons (5244)'. This repetition suggests the incapability to handle the overrepresentation bias, as discussed earlier.

Conversely, the Random Forest and eXtreme Gradient Boosting models manifested precise predictions as anticipated, correctly identifying 'Accounting officer' with 'Accountants (2411)' and 'Foreman' with 'Construction supervisors (3123)'. This level of accuracy suggests these models, while harnessing the power of USE, have a deeper ability to analyse and interpret the broad spectrum of context and details of job titles, going beyond just matching exact keywords to flag the corresponding ISCO codes. Nonetheless, even the most meticulous models, such as eXtreme Gradient Boosting, may face challenges at times when job descriptions are lacking in detail. For example, the model might classify an 'Accounting officer' as 'Accountants (2411)' without considering that the role might potentially fit with 'Accounting Associate Professionals (3313)' or 'Accounting and Bookkeeping Clerks (4311)' as well.

It is important to note here that such complexity of classification also poses big challenges even for human coders. The lack of specification in real-world job title (in comparison to text-book ISCO titles) is a likely factor why we find numerous cases of many-to-many in our original data from DoE in the first place.

The analysis of granular predictions affirms that the eXtreme Gradient Boosting model deserves to be our core classifier for its consistently high accuracy and robust performance across various job classifications, as discussed in length in this section. Its ability to discern fine-grained differences and mitigate the effects of overrepresented classes demonstrate its suitability as the preferred choice for the accurate categorization of varied and complex job titles in the Thai labour market context.

5. CONCLUSION

5.1. Drawbacks

In our exploration of the labour market through advanced NLP techniques, certain constraints of the USE and the nature of our data sources must be acknowledged. While robust and versatile, the USE was originally trained on a diverse set of contexts, which presents a challenge when applied exclusively within a focused domain like the labour market. This general-purpose training may not capture all the subtleties and specialised terminologies unique to the labour market's discourse, in particular, within the Thai context.

Furthermore, the discrepancy between the relatively more formal job postings from the Department of Employment utilised in our training process and the more dynamic, informally styled data scraped from online job platforms introduces another layer of complexity. This variation in data style may affect the model's ability to maintain its high performance in deployment. To address this, our approach includes an additional step that refined the training data to ensure a broad spectrum of writing variations has been recognised by the model.

Additionally, the absence of certain job categories from traditional sectors like agriculture or public-sector jobs, which are rarely present in formal job posts from DOE or job online job advertising, may pose a limitation in analysing the labour market more completely. Similarly, the existing classification of ISCO labels restricts the model's ability to recognise and classify emerging job categories such as green-related jobs. However, despite these challenges, our model demonstrates considerable strength in navigating these limitations, showcasing its capacity to deliver precise classification amidst evolving labour market trends, thereby affirming its utility and versatility in labour market research.

5.2. A Preview of Model Applications

This section highlights the application of the model we developed in this paper, illustrating how the advanced capabilities of our XGBoost model effectively categorise over 1.1 million job posts, unveiling the landscape of labour demand captured from two extensive online job platforms spanning from the fourth quarter of 2020 to the third quarter of 2023. The colossal dataset not only provides a snapshot of current labour market conditions but also enables us to track trends and changes over an extended period, offering stakeholders valuable insights into the dynamics of job availability, the evolution of skill requirements, and the shifts in occupational demand across various sectors. This rich analysis helps to refine workforce strategies, enhance job matching technologies, and shape educational and training programs to meet the real-time needs of the labour market.

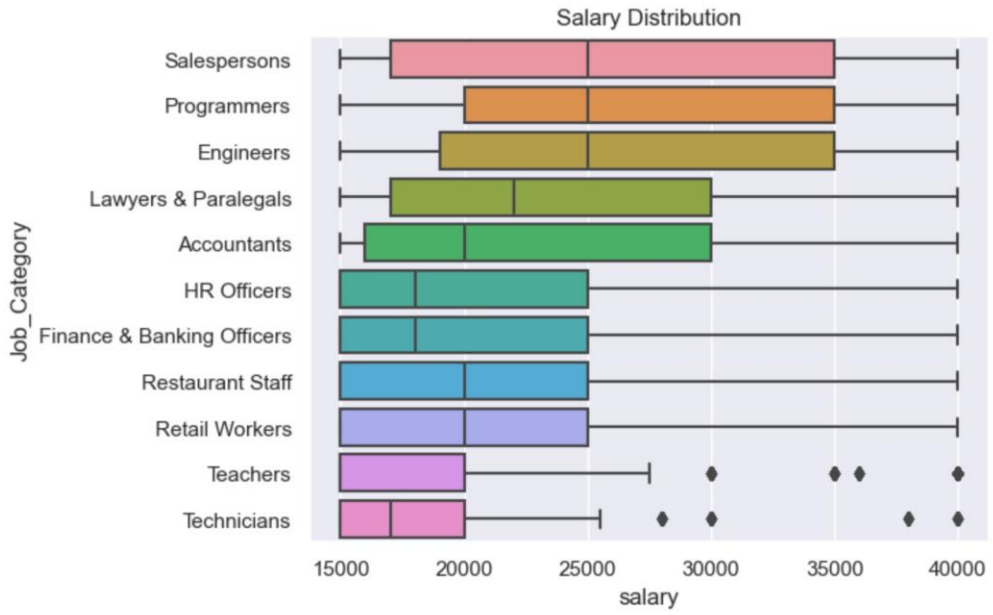

Figure 14: Salary Distribution of Job Advertisement by Job Category

As shown in Figure 14, the salary distribution across major job categories reveals substantial variations. Notably, engineers and programmers exhibit particularly wide ranges and higher end in potential earnings. Engineers show salary ranges extending from approximately 20,000 THB to nearly 40,000 THB, reflecting the diverse skill sets and experience levels within this category. This variability not only highlights the value of specialised skills but also underscores the potential for individuals to elevate their earning capacity through strategic career advancements and skill development. On the other hand, teachers and technicians exhibit a narrower salary range, predominantly spanning from 15,000 to 20,000 baht. This more constrained salary bracket for teachers and technicians suggests a tighter consolidation of roles and responsibilities within these professions, reflecting less variation in the levels of specialization and perhaps more uniform industry standards or regulatory influences impacting pay scales.

We also discover that wage disparities are stark, illustrating by a consistent gender pay gap across all major job categories shown in figure 15. This gap was quantified by calculating the differences in median salaries offered in job advertisements that specified gender preferences—specifically comparing postings that exclusively sought male candidates to those exclusively seeking female candidates. Notably, the gap is particularly pronounced in high-skill roles such as Managers and Programmers, where it can stretch up to 8,000 baht on average, emphasizing that men generally receive higher compensation than their female counterparts in these critical sectors. Most job categories demonstrate a pay gap greater than zero, confirming a systemic issue where males are often compensated more generously across the board. This widespread disparity underscores the urgent need for a comprehensive review and reform of compensation practices within organizations and demands that policymakers intensify their efforts to address and promote gender wage parity.

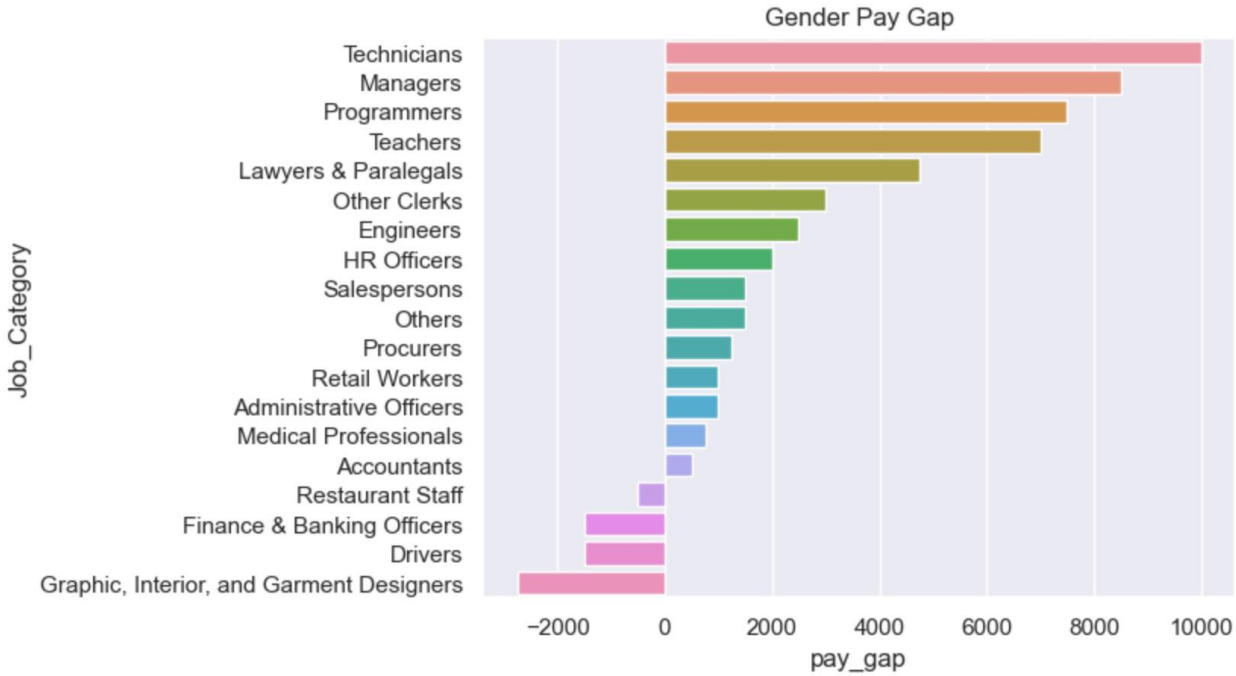

Figure 15: Different of Offered Salary Between Male and Female Candidate by Job Category

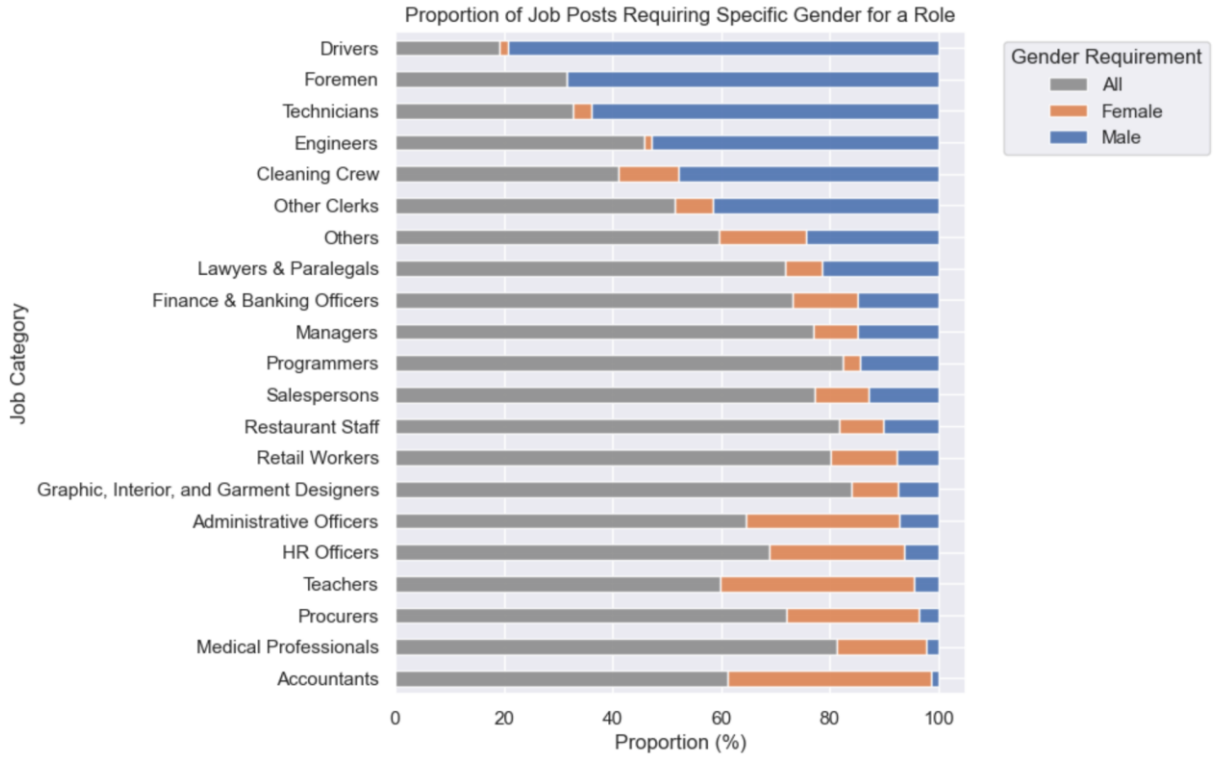

Figure 16: Proportion of Job Requiring Specific Gender for Their Candidates by Job Category

Figure 16, depicting the proportion of job posts requiring specific gender for roles, reveals a striking pattern of gender discrimination in the labour market. The data highlights a substantial bias toward male candidates, particularly in roles traditionally dominated by men such as drivers, foremen, technicians, and engineers. These job categories show a pronounced preference for male applicants, with a significant percentage of job advertisements explicitly specifying a male requirement. This gender specification in job ads not only reflects underlying societal biases but also poses barriers to entry for women in various fields, highlighting a critical area for intervention to promote inclusive hiring practices.

In addition, while the data distinctly shows a prevailing bias towards male candidates in many roles, it also highlights a stereotypical preference for female candidates in roles such as Accountants, medical professionals (notably nurses), and teachers. This pattern underscores traditional gender roles where women are often preferred in caregiving or educational positions, which can perpetuate stereotypes and limit career options for both men and women in these fields.

These findings not only shed light on current labour market conditions but also offer actionable insights for various stakeholders. Educational institutions, for instance, can use this data to align their curricula with real-world salary trends and market demands. Similarly, government agencies and private sector leaders can leverage these insights to formulate strategies that address pay disparities, promote fair hiring practices, and ultimately foster a more equitable labour market. This detailed understanding of job market dynamics is invaluable for driving strategic decisions in workforce development and economic planning.

5.3. Next steps

This paper has laid the groundwork by developing a text classification algorithm that effectively categorizes online job postings into International Standard Classification of Occupations (ISCO) codes, revealing insightful trends and patterns within Thailand's labour market. Leveraging this methodology allows us to harness the extensive data collected from online job portals, transforming raw job listings into structured, ready-to-use information.

The Beveridge Curve, a fundamental concept in labour economics, illustrates the relationship between unemployment rates and job vacancies. It serves as a diagnostic tool, helping to understand the dynamics of job matching in the economy, indicating the efficiency of labour markets, and highlighting shifts caused by economic changes or policy interventions. The curve is particularly valuable for identifying frictions within the labour market, such as the mismatch between the skills of job seekers and the requirements of available jobs.

Building on this foundation, our subsequent, companion paper aims to apply these classifications to map out the Beveridge Curve for Thailand for the first time. This effort will significantly advance the academic and practical understanding of the labour market by providing a real-time analysis of job vacancy and unemployment trends across occupation groups from the information stemming from jobseekers and their resumes on the job websites. By integrating

machine learning techniques with the analysis of online job data, this study will offer a more nuanced view of labour demand and supply, enhancing the granularity with which labour market tightness and skill mismatches are observed.

The academic contribution of this forthcoming paper will be substantial, offering new methodologies and insights into Thai labour market analysis. It will provide a detailed examination of how modern data sources like online job portals can complement traditional economic models and surveys, thereby enriching the empirical framework for labour economics research. This exploration is anticipated to be a valuable resource for policymakers, economists, and scholars, providing actionable insights and a deeper understanding of labour market dynamics in a rapidly changing economic landscape.

REFERENCES

Agrawal, S., & Tiwari, A. (2022). Solving multimodal optimization problems using adaptive differential evolution with archive. *Information Sciences*, *612*, 1024-1044.

Beveridge, William H. (1944). *Full Employment in a Free Society.* New York: W. W. Norton and Company

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*.

Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F., & Lloret, P. (2014). Short text classification using semantic random forest. In *Data Warehousing and Knowledge Discovery: 16th International Conference, DaWaK 2014, Munich, Germany, September 2-4, 2014. Proceedings 16* (pp. 288-299). Springer International Publishing.

Chai, K. E. K., Anthony, S., Coiera, E., & Magrabi, F. (2013). Using statistical text classification to identify health information technology incidents. Journal of the American Medical Informatics Association, 20(5), 980-985.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Da Costa Pereira, C., & Tettamanzi, A. G. B. (2016). Short text classification using semantic random forest. *Proceedings of the 2016 International Conference on Computational Science*.

Deming, D., & Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, *36*(S1), S337-S369.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elsby, M. W., Michaels, R., & Ratner, D. (2015). The Beveridge curve: A survey. *Journal of Economic Literature*, *53*(3), 571-630.

Ghosal, S., & Jain, A. (2023). Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier. *Procedia Computer Science*, *218*, 1631-1639.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). Remote Work across Jobs, Companies, and Space.

He, J., Liu, D., Zhang, C., & Xu, H. (2021). Bert-based multi-label classification with convolutional neural network for emotion classification. *arXiv preprint arXiv:2101.09635*.

Hendrawan, I. R., Utami, E., & Hartanto, A. D. (2022, December). Comparison of Word2vec and Doc2vec methods for text classification of product reviews. In 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 530-534). IEEE.

Hershbein, B., & Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review*, *108*(7), 1737-1772.

Khandve, S. I., Wagh, V. K., Wani, A. D., Joshi, I. M., & Joshi, R. B. (2022, February). Hierarchical neural network approaches for long document classification. In *Proceedings of the 2022 14th International Conference on Machine Learning and Computing* (pp. 115-119).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moreno, M. R., Sánchez-Franco, M. J., & Tienda, M. D. L. S. R. (2023). Examining transaction-specific satisfaction and trust in Airbnb and hotels. An application of BERTopic and Zero-shot text classification. *Tourism & Management Studies*, *19*(2), 21-37.

Nakavachara, V., & Lekfuangfu, W.N. (2018). Predicting the present revisited: the case of Thailand. *Thailand and The World Economy*, *36*(3), 23-46.

Nguyen, T. H., Nguyen, H. H., Ahmadi, Z., Hoang, T.-A., & Doan, T.-N. (2021). On the Impact of Dataset Size: A Twitter Classification Case Study. Association for Computing Machinery.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ramayanti, D., & Salamah, U. (2018). Text classification on dataset of marine and fisheries sciences domain using random forest classifier. *Int. J. Comput. Tech*, *5*(5), 1-7.

Rasooli, M. S., & Tetreault, J. (2018). YaraParser: A fast and accurate dependency parser. *arXiv preprint arXiv:1803.06567*.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sun, D., Xu, J., Wen, H., & Wang, Y. (2020). An optimized random forest model and its generalization ability in landslide susceptibility mapping: application in two areas of Three Gorges Reservoir, China. *Journal of Earth Science*, *31*, 1068-1086.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

Wieting, J., & Kiela, D. (2019). No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning-based natural language processing. *IEEE Computational Intelligence Magazine, 13*(3), 55-75.

APPENDIX

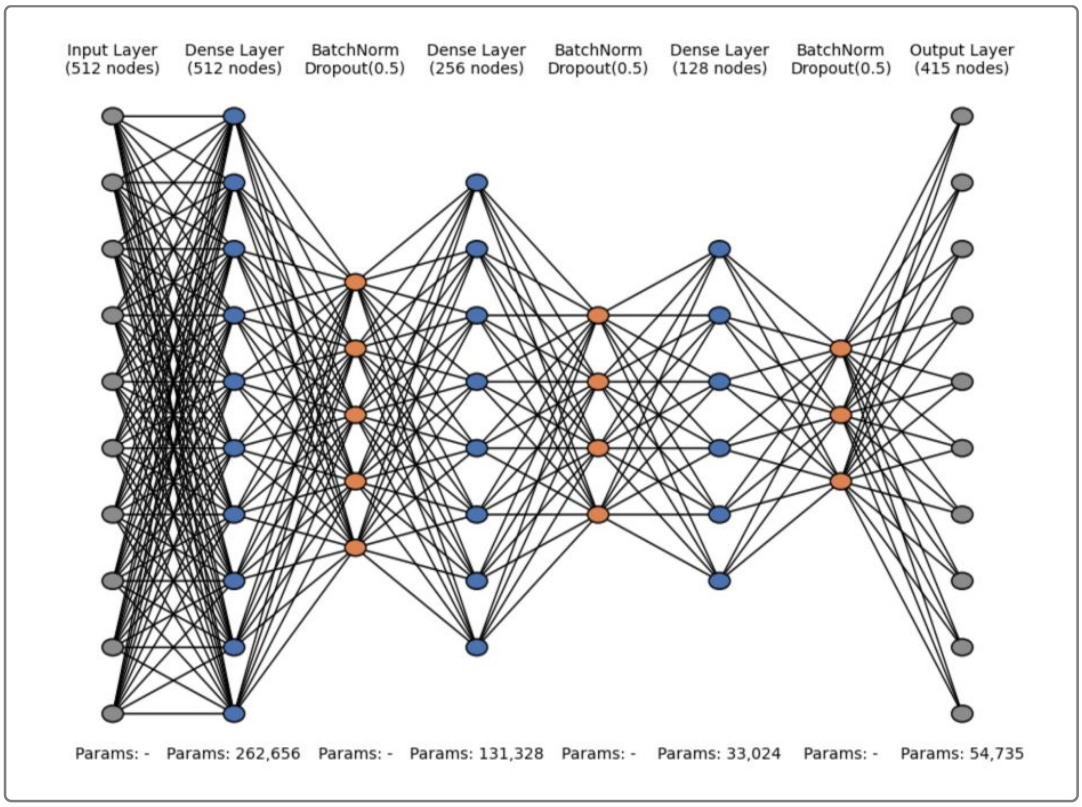

Figure 8: The Neural Network Architecture