applied sciences

MDPI

Article

Age Estimates from Name Characters

Jung-Shiuan Liou [1], Ching-Yen Hsiao [2], Lork-Yee Chow [3], Yen-Hao Huang [3] and Yi-Shin Chen [2,3,*]

[1] International Intercollegiate Ph.D. Program, National Tsing Hua University, Hsinchu 300044, Taiwan; lindaliou000@gmail.com
[2] Department of Computer Science, National Tsing Hua University, Hsinchu 300044, Taiwan; n79122@gmail.com
[3] Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu 300044, Taiwan; chowlorkyee.yvonne@gmail.com (L.-Y.C.); yenhao0218@gmail.com (Y.-H.H.)
* Correspondence: yishin@gmail.com; Tel.: +886-3-571-5131 (ext. 31211)

Abstract: Traditionally, we have been attempting to extract useful features from the massive amount of data generated daily. However, following the legal constraints regarding personal data protection and the challenges of potential data biases and manipulation, artificial intelligence that relies less on big data and more on reasoning ability has become an emerging trend. This paper demonstrates how to estimate age and gender using names only. The proposed two-layer comparative model was trained on Taiwanese names, and its generalizability was further examined on bilingual and cross-border names. By considering additional features of the contextual environment, the model achieves high accuracy in age and gender prediction on Taiwanese and bilingual names. However, the prediction results for ethnic-Chinese Malaysian names (in English) do not reach the same level. This is due to the linguistic differences among Chinese dialects; the features trained on Taiwanese names cannot be directly applied to English names in Malaysia. This study illustrates a path for accomplishing prediction tasks using minimal data and highlights a future possibility for further research.

Keywords: names; age prediction; gender prediction; contextual features; natural language processing (NLP)

Citation: Liou, J.-S.; Hsiao, C.-Y.; Chow, L.-Y.; Huang, Y.-H.; Chen, Y.-S. Age Estimates from Name Characters. *Appl. Sci.* **2021**, *11*, 9611. https://doi.org/10.3390/app11209611

Academic Editor: Stavros Souravlas

Received: 27 September 2021
Accepted: 12 October 2021
Published: 15 October 2021

1. Introduction

In an era when our daily life has become penetrated by the applications of artificial intelligence (AI) and machine learning, data has become the "new dollar". We generate massive amount of data every day in fields like computer vision and natural language process (NLP). Thus, one of the main challenges is how to find and define useful features for identifying objects or distinguishing fake news. Particularly in the domain of NLP, word selections, word relations, and semantic meanings are studied to extract hidden information behind words. For instance, Bidirectional Encoder Representations from Transformers (BERT) explores word sequence to gain contextual meanings of an article [1]. Text classification using Convolutional Neural Network (CNN) extracts features from words themselves [2]. For these approaches, we receive mountains of data as input and try our best to conduct dimensional reduction to acquire useful features.

However, data-hungry methods face legal and ethical constraints. More governments are imposing stringent requirements on the use of personal data. The European Commission issued the General Data Protection Regulation in 2016 to protect personal data and privacy [3]. Similarly, the Chinese government announced the Data Security Law and Personal Data Privacy Law, which came into effect in 2021 [4,5]. Meanwhile, these black-box algorithms receive increasing criticism regarding potential biases and manipulation [6,7]. Therefore, there is an emerging trend that AI will rely less on big data and more on reasoning ability [8]. This means training models to approach problems and tasks using more common sense and ready expertise.

Following the return of ownership of personal data to users' hands, it is becoming harder to acquire detailed data without individuals' consent. More often, what we can access is only names. However, certain demographic attributes like age and gender are crucial to make good recommendations or conduct marketing events. Under the scenario where only minimum data is available, one of the possible ways is to infer results by considering the contextual environment. In this paper, a method to estimate age and gender using names only is demonstrated.

The styles of character-based names from different countries or regions are varied. For example, Taiwanese names mostly consist of three words like "王大明" or "陳淑芬"; Two-character names are more common in China, such as "張浩" or "李燕"; Japanese names usually include more characters like "いとうひろぶみ" or "あべしんぞう". When we study a name, there are not enough inherent features. It is possible to study the last name, first name, characters, or word pronunciation. However, it is insufficient to generate useful results based on these features only. Hence, creating additional features by considering contextual environments, such as culture, society, history, or belief, to extend our feature selections is worthy of study. The corresponding features could be more representative than those directly extracted from names. With the attempt to enlarge our feature basis, we explore various features, such as character combination, pronunciation, fortune-telling elements, and zodiac, to predict the age and gender of a given name.

This paper aims to predict age and gender based on a given name considering local cultural and social contexts. In Taiwan, the naming rules, anthroponymy, usually reflect parents' blessing or expectations of their children. For example, a female name may contain characters or radicals with the meaning of beauty or elegance, such as "美" or "妍". In older generations, the parents may prefer the character "美", which is directly translated as beauty, while modern parents may prefer the character "妍", which implies beauty in a more implicit way. Different generations have their own preferences in choosing characters to represent the same meaning. Furthermore, fortune-telling elements may be considered during the naming process because of Chinese tradition. The five elements (Gold "金", Wood "木", Water "水", Fire "火", and Earth "土"), fortune map "三才五格", and the 12 Chinese zodiacs are popular fortune-telling characteristics. For instance, if a person's fortune map shows a lack of "Water" according to the fortune teller, the parents may give a name with the radical "Water" like "源".

In addition to the above features, how to pronounce a name, the domain of phonology, is also discussed, as we believe it could be an important feature for age and gender prediction. In previous linguistic research, name–gender relations have demonstrated that phonetic characteristics provide evidence about identifying the gender of a name [9]. The pronunciation of a name contains vowels, consonants, and tones (though there are some characters pronounced as soft tone, such as "們" or "呢", they are not applied to names). We further attempt to apply phonetic characteristics for age prediction. Finally, a cross-border experiment using ethnic-Chinese Malaysian names is conducted to examine the generalizability of the model. The major contributions of this paper are listed below.

- Create additional features beyond a given name, such as fortune-telling elements, Chinese zodiac, and word radical, from the contextual environment.
- Explore how the name features impact the prediction results of age and gender estimation.
- Examine the generalizability of the model trained by Taiwanese names on bilingual names (Chinese/English names) and ethnic-Chinese Malaysian names.
- Present the methods of preprocessing name data from multiple sources in Taiwan and Malaysia.

2. Related Work

In this section, the methods of learning semantic meanings from words are examined. Then, what demographic characteristics have been explored from names is studied. Finally,

the approaches previously applied to age estimation are discussed. Learning from these past works, the design of our proposed model is illustrated in the following section.

2.1. Semantic Learning from Words

In deep learning approaches, words are embedded as vector representations using the learning techniques such as Word2Vec and GloVE [10,11]. The word vectors retain the semantic reasoning capability and are utilized as the input feature for deep learning models that are designed to capture the semantic meaning of sentences and documents. The widely applied context-based deep learning models are sequence-based models, Recurrent Neural Networks (RNN), and large-size transformer-based models with multi-head self-attention structures. The RNN-based models are commonly adopted in the tasks such as learning sentence or document representations by focusing on the word sequence order [12–14]. Later, Bahdanau et al. proposed the attention mechanism that dynamically gives weight for each word on a sentence level or on a document level [15]. Recently, following the trend of learning pretrained language models, several transformer-based models, such as BERT and XLNET, were developed with multi-head attention and masked token training mechanism [1,16,17]. The transformer-based models currently achieve the state-of-art performance on most NLP tasks.

2.2. Demographic Characteristics from Names

Previous works explored demographic characteristics using names from different sources or combined with other factors. Several works predicted gender based on users' names only [18,19]. Burger et al. distinguished gender on Twitter using names and screen names for classification [20]. Brown proposed a naive Bayes classifier to classify the gender of the name using simple features from a NLTK book [21]. Hu et al. proposed a linear model combined with judicious feature engineering to predict gender and achieved the same level performance as using character-based Long Short-Term Memory (LSTM) or BERT [22]. Other works studied the structure of given names and how they are associated with gender and ethnicity [23,24]. One study conducted a large-scale machine learning framework to predict ethnicity using a novel set of personal name and census location data [25]. Lee et al. predicted ethnicity and nationality based on RNN models using the names on Olympic record data [26].

2.3. Age Estimation

In terms of age estimation, text analysis and facial recognition are the two approaches mainly applied. Text analysis is widely used to capture features of writing styles from articles or blogs on social media to estimate authors' ages [27–30]. The feature selection includes part-of-speech (POS) features, psychological word features from linguistic inquiry and word counts (LIWC), and punctuation usage [31]. Additionally, developing age and gender predictive lexica is a method commonly used on social media [32,33]. Conventional methods of facial recognition utilize facial images to predict ages [34–37]. Due to the complexity of this task, some works try to capture facial features by dividing a facial image into local parts or using hierarchical deep-learning classifiers [38–41]. Abousaleh et al. compared the relations between the reference images of known ages with an unknown facial image to aggregate the hints to conduct age estimation [42].

3. Materials and Methods

The goals of this paper are to predict the gender and age interval of an individual using Taiwanese names and to verify the generalizability of the proposed framework on bilingual names and ethnic-Chinese Malaysian names. Meanwhile, the names used in this research are focused on their given names. For the Taiwanese name dataset, the problem is addressed as:

Definition 1. *Given a Taiwanese name $N = \{c\}$, $c \in C$, where C is a set of Chinese characters; N can be presented as $N^{(first)} \cup N^{(last)}$ where $N^{(first)} = \{c|c \in C\}$, $N^{(last)} = \{c|c \in C\}$, $N^{(last)} \in NL$ and NL is a set of Taiwanese family names. We denote a model f to estimate the age interval $\hat{y} \in \{[y, y+5) \mid y \in \{1940, 1945, ..., 2010\}\}$ of an input name N, such that $\hat{y} = f(N)$.*

The basic features of this problem are character-based features that can reflect the meanings of the name itself and then help with understanding the popular characters used in different generations. By precisely capturing these features, the results of age estimation can reach high accuracy.

Naming customs vary among different ethnic groups. Therefore, the whole task was divided into two phases. In the first phase, we constructed features based on given names and then built an age-interval classifier. In the second phase, to examine the generalizability of the model, the Taiwanese Public Figure Name dataset and the Ethnic-Chinese Malaysian Name dataset were applied to conduct cross-border tests.

According to the statistics of the collected data, the range of birth year was from 1940 to 2010. Hence, the output age values ranged from 1940 to 2010. An overall framework is presented in Figure 1.

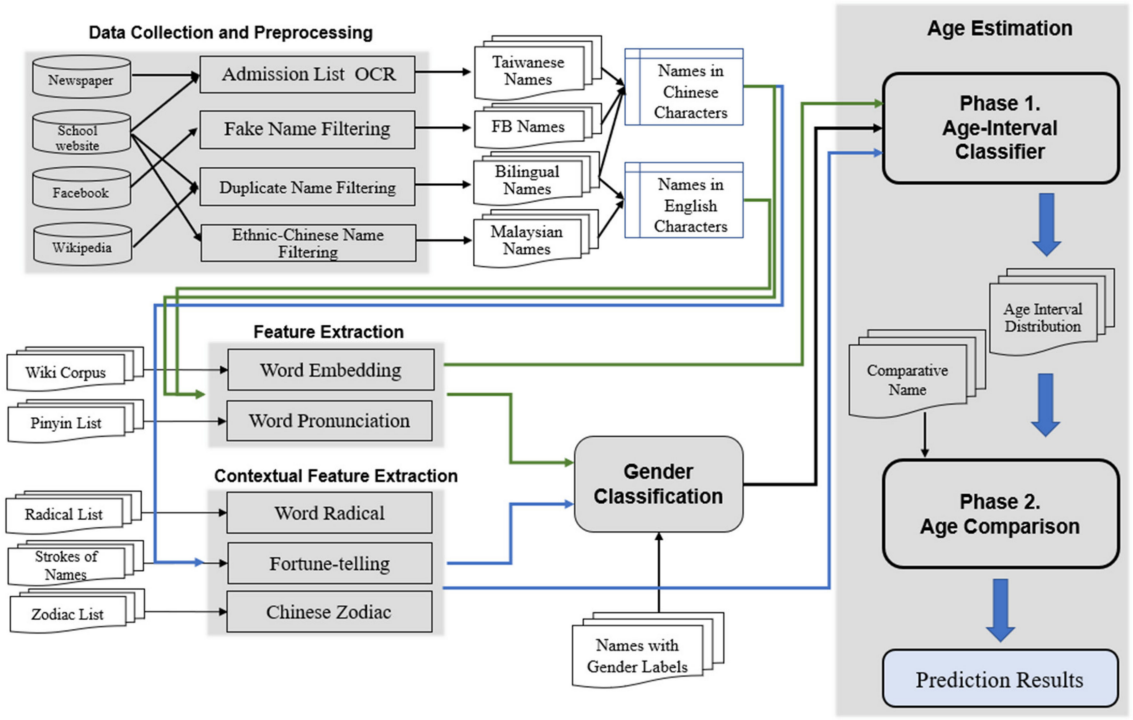

Figure 1. Overall framework.

3.1. Data Collection

Taiwanese names are based on Chinese characters. In this paper, four datasets were collected from multiple sources: the Taiwanese Real Name dataset, the Taiwanese Social Media Name dataset, the Taiwanese Public Figure Name dataset, and the Ethnic-Chinese Malaysian Name dataset. The former two datasets were collected from our previous work [43]. The details of the collection process are elaborated below.

3.1.1. Taiwanese Real Name Dataset

Most Taiwanese people follow the same path for receiving education—sitting an entrance examination before each phase of school entrance. The examinations were held by the Taiwanese government Joint College Entrance Examination from 1954 to 2002. The names of students who passed the examinations were publicly released in various formats

over time. More than 100,000 high school graduates took the examination each year. To collect as many names as possible, the real-world names were collected in three ways.

(A). Newspaper Optical Character Recognition (OCR): In the early stage, the Taiwanese government published the student names along with their qualified schools in newspapers. The scanned photos of the lists were downloaded, and the names were recognized by *ABBYY FineReader* software, which performs best for old newspapers. (B). College Admission List: The electronic versions of the admission lists have been available since 1994. The files of Taiwanese names on the admission lists from 1994 to 2012 were acquired from the website of Professor Tsai [44]. (C). Graduation Name List: To obtain extended coverage of the names collected, the graduation name lists from reunion websites were gathered. The data were manually collected using suitable keywords on the search engine.

3.1.2. Taiwanese Social Media Name Dataset

Additionally, Taiwanese names from the popular social media website, Facebook (FB), were collected. Although the registration process for FB does not request a real name, users tend to have a username close to their real one. Meanwhile, there are several fanpages that provide fortune-telling games that attract many Taiwanese users. Users provided their birth dates as input and received replies, such as "Excellent," regarding their fortune of love. The data was included as the FB dataset.

3.1.3. Taiwanese Public Figure Name Dataset (Bilingual Names)

To explore the influence of name features on age-interval prediction regarding a person who has a Chinese name and an English name simultaneously, the bilingual names of Taiwanese public figures from Wikipedia and Taiwan local university faculty websites were collected. The process was manually collected using keywords, such as "Taiwanese actors," "Taiwanese athletics," and "Taiwanese," and obtaining professors' names from official websites. The data includes their Chinese name, English name, gender, and year of birth. As a result, a total of 1002 Taiwanese public figure names were collected for the years 1900–2009.

3.1.4. Ethnic-Chinese Malaysian Name Dataset

Due to the privacy policy, name data in Malaysia can only be collected from the university graduation lists. We downloaded the data from the Multimedia University graduation portal from 2000 to 2018, comprising 19 graduation sessions in total. After combining all the names and removing those with non-Chinese name formats, 30,715 names were collected. By looking at a person's degree, their age can be easily estimated using Equation (1) and Table 1, as shown below.

$$BirthYear = intakeyear - typicalage \quad (1)$$

Table 1. Malaysian education system.

Level/Grade	Years of Education	Typical Age
Primary school	6	7–12
Secondary school	5	13–17
Pre-university	2	18–19
Bachelor of degree	3–4	20–23/24
Master's degree	2	25–26
Doctorate degree	Vary	Vary

The overview of the datasets is shown in Table 2. As the collected names were not evenly distributed among different datasets and age intervals, the data size was re-sampled in the following stages.

Table 2. Dataset overview.

Dataset	# of Names	Year Range
Taiwanese Real Name	1,635,533	1940–2000
Taiwanese Social Media Name	731,137	1940–2010
Taiwanese Public Figure Name	1002	1900–2009
Ethnic-Chinese Malaysian Name	30,715	1972–1998

3.2. Data Preprocessing

As our datasets were collected from multiple sources in Taiwan and Malaysia, the regular name input was prepared for future manipulation. The proposed method is two-step data preprocessing for each dataset.

3.2.1. Name Filtering

(A). Taiwanese Name Dataset

Taiwanese names are based on Chinese characters, basically two to four characters. Thus, the names with non-Chinese characters and those that have more than four characters were dropped. For Taiwanese real names, duplicate names on school lists were excluded, as they may be the same person in different grades. For names collected by the OCR process, the simplified Chinese characters were switched into the traditional Chinese characters used in Taiwan. For Taiwanese social media names, duplicate names and IDs were excluded to avoid the same person playing fortune-telling games multiple times.

(B). Malaysian Name Dataset

Malaysia is a multi-ethnic country, with people originating from China, Taiwan, Malay, India, etc. Based on the study on Malaysian naming culture, we manually labeled the Malaysian name dataset by classifying ethnic groups. The formats of names from different ethnic groups vary. For Malay, the way of naming consists of the word "BIN" or "BINT"; for Indian, "A/L" or "A/P" is shown in their names; as for Chinese, the names usually contain no more than four words. As our study is focused on ethnic Chinese, all the names with non-Chinese name formats were removed. An example of labeling ethnic groups is presented in Table 3.

Table 3. Labeling ethnic groups.

No	Name	Ethnic Group
1	HAFIZUDIN BIN ABU HANIFAH	Malay
2	CHOW LORK YEE	Chinese
3	GURD SINGH A/L NACAR SINGH	Indian

3.2.2. Family Name Segmentation

(A). Taiwanese Name Dataset

According to the latest statistics from the Ministry of Interior in Taiwan, there are 1832 family names. The top 100 family names account for 96.57% of the total population of Taiwan. We only retained the names in which the family name $N^{(Last)}$ belongs to the top 1500 family names. Most Taiwanese family names are a single Chinese character, such as "陳" (Chen). However, some of them combine two Chinese characters, such as "張簡" (Zhang Jian). Some people add another's family name in front of theirs because of marriage. Although the probability of having a family name with two Chinese characters is low, our research took this scenario into account.

(B). Malaysian Name Dataset

As different countries have different sequences of first and family names, the naming style should be learned first. For the Taiwanese style, the family name is arranged after

the first name. Conversely, for the Malaysian style, the family name is arranged before the first name. By identifying the styles, family name segmentation can be conducted for the English names of Taiwanese and Malaysian names. An example of English family name segmentation is shown in Table 4.

Table 4. English family name segmentation.

Name	Family Name	Style
Lork-Yee Chow	Chow	Taiwanese style
Chow Lork Yee	Chow	Malaysian style

3.3. Feature Extraction

In Taiwan, when parents name their child, they may consider their expectations and blessings, fortune-telling elements, pronunciation, social factors, etc. The characters chosen for different genders might vary as well. The features below are extracted from the name itself or combined with contextual considerations when giving a name. To study names, knowledge of anthroponymy and phonology are applied.

3.3.1. Word Meaning (W)

The name characters represent different meanings, which usually link to parents' wishes. With the different age generations and genders, parents will choose different words for their children. For example, parents in the early generations may give the name "添財", which means "bringing fortune" under the background of living in a rural area. However, such a name is rare in recent years, as the standards of living have been lifted following economic growth.

Usually the meaning of a word depends on which key phrase it is linked to. Taking a word "圓" (circle) as an example, it can be used as "圓滿" referring to being perfect, or "圓潤" referring to being mellowed. However, what is critical for naming is capturing the meanings of the single words used in names. The similarity among words is more important. Therefore, the purpose of training is to find similar words of the same meaning. Different from the transformer-based models such as BERT and XLNET which explore multiple meanings of the same word used in different contexts [17], the *Word2Vec* model with the focus of capturing the word meanings was chosen here [9].

Then, the *Word2Vec* model was trained with different settings, including oral usage, example sentences on the dictionary, poem, and Wikipedia sentences, shown as Figure 2. The results demonstrated that the combination of example sentences on the dictionary and Wikipedia performed best and therefore was decided to be the word embedding model. The reason is because training on example sentences can do better in capturing the similarity of words than training on other data.

3.3.2. Fortune-Telling Feature (F)

In Taiwanese culture, fortune telling is the ancient wisdom to infer the future of individuals [45]. There was a period when people might refer to the suggestions, in terms of lucky or unlucky, on the Chinese Fortune Calendar when choosing a day for moving or traveling [46]. Still now, people may ask for help from a fortune teller to give a suitable name. People tend to believe that a good name might be helpful to the fortune of their children, and there is no harm to include fortune-telling during naming. Fortune tellers utilize varying techniques, such as palm and face reading "手相和面相", the four pillars: hour, day, month, and year of birth "生辰八字", or name strokes "姓名筆劃" to predict personality, marriage, occupations, and future incidents. Therefore, Taiwanese naming rules with fortune-telling elements, anthroponymy, have become part of the tradition [47].

Figure 2. *Word2Vec* with different settings: (**A**) oral usage, (**B**) example sentences in the dictionary, (**C**) poem, (**D**) example sentences in the dictionary and Wikipedia.

There are two primary rules used in Taiwanese anthroponymy: a fortune map "三才五格" and a Chinese zodiac "生肖". In the fortune map, a name is originated from three basics ("三才"-天格, 地格, 人格) and extended to five elements ("五格"-天格, 地格, 人格, 外格, 總格). "天格" refers to the family name, which is related to fortune during childhood. "地格" refers to the first name, which is related to fortune during adolescence. "人格", as the core of a name, refers to the first character of the name, which can affect the whole life. "外格" is related to fortune during middle age. "總格" is related to fortune during old age. As a total, the fortune of an individual is completely presented through a fortune map. The rule of a fortune map, which calculates the strokes of Chinese characters of a name, is illustrated in Table 5. The Chinese zodiac will be discussed in the next section.

Table 5. The rule of the fortune map.

Structure	Type	Rule
天格	單姓	Stroke ($N^{(Last)}$) + 1
天格	複姓	Stroke ($N^{(Last)}$)
人格	單姓	Stroke ($N^{(Last)} + 1_{st}N^{(First)}$)
人格	複姓	Stroke ($2_{nd}N^{(Last)} + 1_{st}N^{(First)}$)
地格	單字	Stroke ($N^{(First)}$) + 1
地格	複字	Stroke ($N^{(First)}$)
外格	單姓單名	2
外格	單姓複名	Stroke ($2_{nd}N^{(First)}$) + 1
外格	複姓單名	Stroke ($1_{st}N^{(Last)}$) + 1
外格	複姓複名	Stroke ($1_{st}N^{(Last)} + 2_{nd}N^{(First)}$)
總格	N/A	Stroke ($N^{(Last)} + N^{(First)}$)

3.3.3. Chinese Zodiac (Z)

The Chinese zodiac is a classification method that assigns a representative animal, namely Rat, Ox, Tiger, Rabbit, Dragon, Snake, Horse, Goat, Monkey, Rooster, Dog, and Pig, for each year in 12 consecutive years. The 12-year cycle continuously repeats itself. Each animal has its own unique characteristics, which are important references for Taiwanese anthroponymy. Some parents may hope their child is born in a specific year of the zodiac, especially the year of the Dragon. According to Chinese culture, being born in this year is considered auspicious [48].

As the Chinese zodiac repeats every 12 years, the order of the Chinese zodiac was utilized as a feature. Based on the age interval of five years, there are 12 combinations from 1945 to 2004, as shown in Table 6. When the model predicts the age of a Taiwanese name, the feature of the Chinese zodiac can enhance the accuracy. For example, according to the model, an age estimation is possibly in either interval A (Rooster, Dog, Pig, Rat, and Ox) or interval B (Tiger, Rabbit, Dragon, Snake, and Horse) with equal probability. Under this circumstance, if a Chinese zodiac characteristic of Pig can be identified in the name, the model may classify the name into interval A.

Table 6. The 12 combinations of the Chinese Zodiac, 1945–2004.

1945–1949	1950–1954	1955–1959	1960–1964
Rooster, Dog, Pig, Rat, Ox	Tiger, Rabbit, Dragon, Snake, Horse	Goat, Monkey, Rooster, Dog, Pig	Rat, Ox, Tiger, Rabbit, Dragon
1965–1969	1970–1974	1975–1979	1980–1984
Snake, Horse, Goat, Monkey, Rooster	Dog, Pig, Rat, Ox, Tiger	Rabbit, Dragon, Snake, Horse, Goat	Monkey, Rooster, Dog, Pig, Rat
1985–1989	1990–1994	1995–1999	2000–2004
Ox, Tiger, Rabbit, Dragon, Snake	Horse, Goat, Monkey, Rooster, Dog	Pig, Rat, Ox, Tiger, Rabbit	Dragon, Snake, Horse, Goat, Monkey

3.3.4. Word Radical (R)

A Chinese radical is a graphical component of a Chinese character under which the character is listed in the dictionary. The radical, as an important component in Chinese fortune-telling, is counted as the number of strokes of a Chinese character. Additionally, some families require different radicals for the names of different generations. The most commonly accepted radicals for traditional Chinese characters consist of 214 entries. Table 7 below is an example of a traditional Chinese radical table. One-hot encoding was used to express this categorical information.

Table 7. Example of traditional Chinese radical table.

Radical	English	Stroke	Character Used
艸	Grass	6	草, 花, 茶
水	Water	4	海, 江, 河
木	Wood	4	樹, 林, 本
手	Hand	4	推, 拉, 提
口	Mouth	3	呢, 只, 古
心	Heart	4	怪, 憶, 忘
虫	Insect	6	獨, 蝴蝶, 螞蟻
竹	Bamboo	6	籠, 笑, 答
言	Speech	7	說, 詞, 語
糸	Silk	6	紙, 給, 紅

3.3.5. Word Pronunciation (P)

(A). Taiwanese Name

The pronunciation of each word of the given name was utilized as a feature. Hanyu Pinyin, often abbreviated to Pinyin, is a system of phonetic transcriptions of standard Chinese in Taiwan. The system includes four diacritics denoting tones, as shown in Table 8. Pinyin without tone marks is used to spell Chinese words in Latin alphabets. Unlike European languages, which consist of clusters of letters, the fundamental elements of Pinyin are initials and finals. There are 23 initials and 33 finals, as indicated in Table 9. Each Chinese character has its corresponding Pinyin, exactly one initial followed by one final. Based on the rules of Pinyin, we converted Chinese characters into Pinyin and further separated Pinyin into initials, finals, and tones. An example of Pinyin rules utilized on Chinese characters is depicted in Table 10.

Table 8. Four tones in Pinyin.

Tone	Symbol
First	—
Second	/
Third	v
Forth	\

Table 9. Elements of Pinyin.

Type	List
Initial	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w
Final	a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, ue, ui, un, uo, uu/v

Table 10. Example of Pinyin rules on Chinese characters.

Character	Syllable	Initial	Final	Tone
明	Mi'ng	M	ing	2
文	We'n	W	en	2
曉	Hsia^{v}o	Hs	iao	3

(B). English Name

When a Taiwanese name is given, usually there is a corresponding English name translated from the pronunciation of the Taiwanese name. The tones of Chinese characters are neglected during the translation process. For instance, the names "偉成" and "薇澄" are both pronounced as "Wei-Cheng" but in different characters. Therefore, it is common for different Chinese characters to have similar pronunciations in English. Meanwhile, the

features, such as fortune-telling or Chinese zodiac, cannot be applied to English names for making predictions. However, even with such differences, it is interesting to learn whether the English name directly translated from the Taiwanese name can be used to estimate an individual's age.

Like many languages, English has wide variation in pronunciation, both from historical development and from dialect to dialect. Here, we indicated the pronunciation of English words with the International Phonetic Alphabet (IPA). This is a system based on phonetic notation of Latin alphabets designed by the International Phonetic Association to be the standardized representation of the sound of spoken language. The main advantage of using the IPA instead of other forms is that we can apply it to multiple languages in the future without significantly changing the format. Figures 3 and 4 show the IPA symbols for consonants and vowels separately [49].

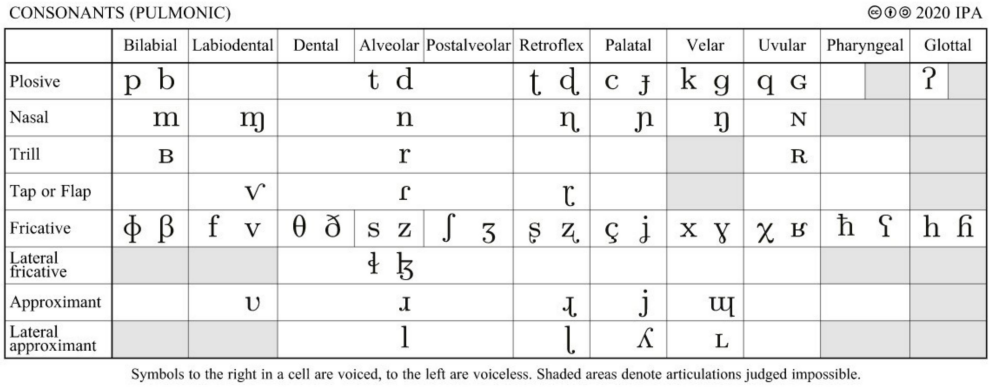

Figure 3. IPA consonants.

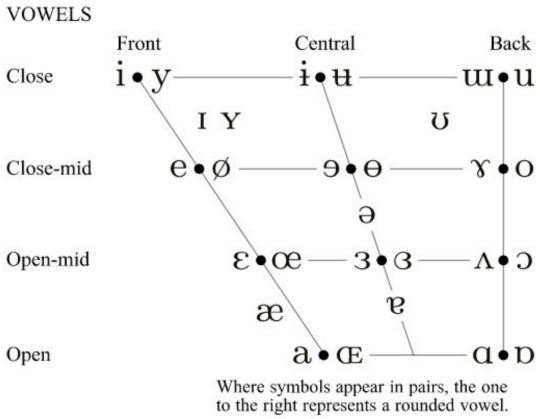

Figure 4. IPA vowels.

Although the IPA is commonly recognized to represent the pronunciation of English words, some names still failed to be converted when we were working on our datasets. This failure stemmed from the differences of various dialects, such as Mandarin, Hokkien, Cantonese, Teochew, etc. The IPA cannot recognize dialect, so those words are unconvertable. Considering this situation, we then looked at the consonants and vowels to get the pronunciation of names.

The modern English alphabet is a Latin alphabet consisting of 26 letters. The letters A, E, I, O, and U are considered vowel letters, while the remaining letters are consonant letters. According to this rule, the English names, directly translated from their Taiwanese names, were separated using Algorithm 1. An example of separating an English name is shown in Table 11. According to the principle of Chinese pronunciation, a word "Lork" consists of only one consonant "L" and one vowel "ork". Though "r, k" does not belong to vowels, there is no further separation for names.

Algorithm 1. Syllable separation of English names

```
Given: V = [A, E, I, O, U, a, e, I, o, u]
con = [ ]
vow = [ ]
for wordlen in word and split into each letter do
    length ← length of word
    for i in the range of length do
        current ← wordlen[i]
        if current is not V then
            add current into con
            move current to next length
        else if current is V then
            split current
            vow ← the rest of the wordlen
        end if
    end for
end for
```

Table 11. Example of separating an English name.

Word	Consonant	Vowel
Chow	Ch	ow
Lork	L	ork
Yee	Y	ee

3.3.6. Gender Probability of Name (G)

A classifier was trained to discriminate the gender of a name with the word meaning, word radical, word pronunciation, fortune-telling feature, etc. An ensemble supervised machine-learning approach, the Random Forest Classifier (RFC), was empirically selected to be our training model. The labeled data was acquired in two ways: the names of single-gender education schools and FB names with gender labels. In Taiwan, there are several single-gender schools in which all the students are either male or female. FB names were collected through the fortune-telling games on the website, in which users need to input gender and birthday (such as M19850312, "M" is male, "19850312" is the birthday) and receive a reply of their love fortune. The model was trained using the combination of the two types of labeled data. The output of gender probability was used as a feature.

3.4. Model Design

The model adopted additional features and was trained based on the prior framework to improve the performance [43]. Meanwhile, it was further applied to bilingual names and ethnic-Chinese Malaysian names. First, all the name features mentioned above—including word meaning, fortune-telling feature, Chinese zodiac, word radical, and word pronunciation—were utilized to classify the gender of a given name. Second, the gender classification result together with other features were used as input to generate probability distribution in each age interval.

In practice, humans have several judgments to estimate the age generation of a given name, but they may have difficulty in giving precise estimation without any assisted method or tool. For instance, we may judge the name "淑芬" as popular in early generations. However, it is difficult to ascertain whether the name belongs to the 50s, 60s, or even older age groups. If reference information with the known age is provided, people can compare and make judgments about the age prediction more precisely. Following this concept, a two-layer comparative estimation model is proposed, including an age-interval classifier and an age comparison mechanism, as illustrated in Figure 5. In the first layer, an individual's age is regarded as a multi-label question and corresponding multiple answers. In the

second layer, inspired by Abousaleh et al.'s work, an estimated age interval is selected after multiple comparisons [42].

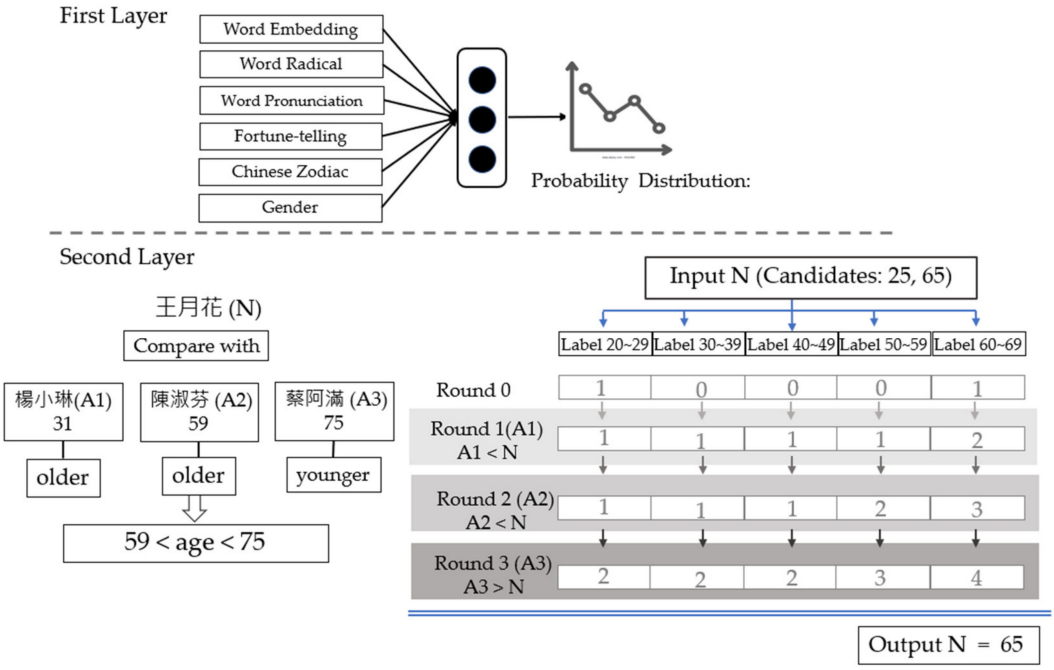

Figure 5. Mechanism of two-layer comparative estimation model.

To speed up the comparison time and minimize the comparison cycles, experiments were conducted to investigate the optimized comparative name size and initial point using the RFC. For each age interval, the best size was 100 names. Additionally, starting from the middle point can effectively reduce the required comparison cycles.

Figure 5 illustrates how the two-layer comparative estimation model works. In the first layer, with the designed features of name N, the neural network outputs the probability distribution of each age interval. The comparative mechanism is demonstrated in the second layer. The right side provides the details of the comparison process. The above example shows that, after the first layer, the two possible candidates for age estimation of the input N are 25 and 65 years old. To further help with precise prediction, three reference names with known ages were selected, namely "楊小琳" (A1) in 31 years old, "陳淑芬" (A2) in 59 years old, and "蔡阿滿" (A3) in 75 years old. Meanwhile, though there should be many age intervals in this example, a simple way that five labels were used to represent age intervals of 20~29, 30~39, 40~49, 50~59, and 60~69 is illustrated here. The comparison process follows the following steps: in Round 0, both Label 20~29 and 60~69 receive 1 point according to the initial result from probability distribution; next, in Round 1, N is estimated to be older than A1, so 1 point is added to Labels 30~39, 40~49, 50~59, and 60~69. Following in Round 2, N is again estimated to be older than A2, therefore, Labels 50~59 and 60~69 get 1 point. Finally, in Round 3, N is estimated to be younger than A3, all the labels receive an additional 1 point. As a result, Label 60~69 accumulates the highest 4 points. Hence, the age estimation of the name N is 65 years old.

As for the bilingual names, the model trained on Taiwanese names was applied. However, due to linguistic differences, the name features that can be extracted from English names are more limited compared to those from Chinese names. Only word embedding and word pronunciation can be utilized as the features for English names.

To further examine the cross-border learning generalizability of the model, the dataset from a different country was selected: ethnic-Chinese Malaysian names, expressed in English. In Malaysia, ethnic Chinese may speak in different dialects, such as Cantonese, Hokkien, Teochew, or Chinese. Each dialect has its own unique pronunciation. The naming

rules might be affected by the dialect used by a family. In this experiment, Malaysian names were used as the testing dataset. Like the previous process, an age interval classifier was first established using the RFC, which generated the probability distribution of each age interval. Then, the target name was repeatedly compared with a reference name with the known age until the target name was in the same age interval as the reference name.

4. Experiment and Results

4.1. Experimental Setup

The model was trained on Taiwanese names, and then the generalizability was verified on bilingual names and ethnic-Chinese Malaysian names. For Taiwanese real names, OCR was used to scan the downloaded images from college admission lists in newspapers from 1958 to 1994. The newspapers from 1988 to 1992 were missing; therefore, a total of 31 years of Taiwanese names were collected. Meanwhile, the electronic versions of the college admission lists were available from 1994 to 2012, in which the file of 2009 was missing. Additionally, student names from over 50 school yearbooks were acquired. For Taiwanese social media names, FB names and all the posts of fortune-telling games were collected until September 2017. In total, there were 4.04 million messages from 48 fanpages. The statistics of the data are presented in Tables 12 and 13.

Table 12. Collected Taiwanese datasets.

Dataset	# of Names	Unique Names	Year Range
Taiwanese Real Name	1,635,533	280,274	1940–2000
Taiwanese Social Media Name	731,137	173,209	1940–2010

Table 13. Collected Taiwanese datasets with gender labels.

Dataset	# of Names	Male	Female
Taiwanese Real Name	77,097	48,508	28,589
Taiwanese Social Media Name	68,826	36,428	32,398

Here, two observations can be identified regarding the two datasets. First, people tend to use fake names or nicknames instead of real names on social media. Even with a fake-name filtering step on social media names, stacked first names (such as "莉莉") and single-character first names (such as "建") appear more often compared to real names. According to the statistics, the ratio of stacked first names on social media is 3.4%, compared to 0.27% on real names; while the ratio of single-character first names is 8.14%, compared to 1.4% on real names. Second, the age distributions of the two datasets are different, as indicated in Figure 6, which refers to the unbalanced data issue.

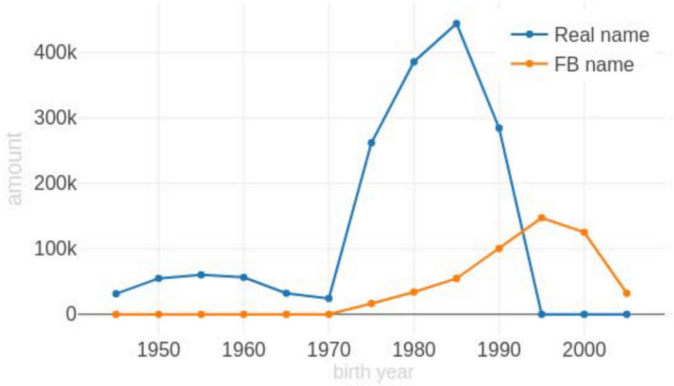

Figure 6. Distribution of names of two datasets.

In the following experiments, each age interval was set of five years, starting with 1945. To solve the data imbalance issue, the continuous years with sufficient data sizes were selected such that balanced data was reached for each age interval, shown as Table 14.

Table 14. Statistics of the sampled datasets.

Dataset	# of Names	Unique Names	Birth Year	# of Intervals
Taiwanese Real Name	302,430	93,522	1945–1995	10
Taiwanese Social Media Name	152,782	57,869	1975–2010	7

For Malaysian names collected from the university graduation lists, the age may be varied for a person with a doctorate degree. Thus, the names with "Doctor of" were removed from the Malaysian name dataset. Additionally, the same sampling approach was adopted to generate a balanced dataset. Only experiments for age-interval prediction were conducted because the gender label is not available for Malaysian names.

4.2. Evaluation Method

To evaluate the performance of the proposed model, the experiments were designed in three directions. The first direction compared the performance using different features and feature combinations for gender prediction and age-interval prediction. Through a series of experiments, how these features impacted the prediction results can be elaborated. The second way examined the generalizability using bilingual names—Chinese names and their corresponding English names. Finally, a different country dataset, Malaysian names, was applied to understand how our proposed model and features support cross-border situations.

In this research, different methods were adopted to evaluate the performance of gender and age-interval prediction. Gender prediction was treated as a binary classification task; therefore, prediction accuracy and the F1 score were applied. The age-interval prediction was regarded as a multi-label classification task with multiple answers. Thus, the metrics of average year error and multi-answer accuracy were utilized to evaluate the performance. The year error calculates the gap between the predicted age interval and the exact birth year of the given name. For example, the model predicts that the age interval of "淑芬" is 1950–1954, and the birth year of this name is 1955. Then, the year error is considered as 1. The smaller the average year error, the better the performance of the proposed model. In this model, the minimum average year error is more important than higher multi-answer accuracy.

The data was divided into a training set and a testing set at 70% and 30%, respectively. The RFC and Multi-Layer Perceptron (MLP) were empirically selected as our classifiers during the model training. In this part, different settings used for uni-class classification of the age-interval classifier were checked, shown as Table 15. The combination of RFC and MLP had the best performance. The two datasets (the Taiwanese Public Figure Name dataset and the Ethnic-Chinese Malaysian Name dataset) used to examine the generalizability of the model were excluded from the training and testing datasets.

Table 15. Different settings of the comparative framework.

Age-Interval Classifier	Comparative Classifier	Real Name		FB Name	
		Year Error	Multi-Ans. Acc.	Year Error	Multi-Ans. Acc.
RFC	RFC	4.11	0.468	3.76	0.407
RFC	MLP	3.90	0.481	3.61	0.428
MLP	RFC	4.10	0.465	4.13	0.385
MLP	MLP	4.22	0.463	3.94	0.394

4.3. Experimental Results

To demonstrate how the designed features can improve the prediction results and generalizability of the proposed model, the performance of the trained model will be discussed. Following this, the prediction results of the extended datasets will be examined.

4.3.1. Performance on Taiwanese Names

The overall performance of gender classification is demonstrated in Table 16. It shows that even with the basic uni-gram character feature (U) or word embedding feature (W), the model can distinguish gender with high accuracy. Between these two, word embedding is more informative than uni-gram in terms of gender classification. To enhance the accuracy, features like first character (FC), second character (SC), word pronunciation (P), fortune-telling feature (F), zodiac (Z), and radical (R) were considered. All the additional features together can slightly improve the accuracy by 0.7%. Meanwhile, the second character was found to be better than the first character for identifying gender.

Table 16. Gender classification accuracy on Taiwanese names.

Feature	Accuracy
FC_P	0.8225
SC_P	0.8612
U	0.8962
W	0.9394
WU	0.9433
WFZRU	0.9446
WPU	0.9458
WZRU	0.9464

As for the age-interval prediction, similarly, word embedding (W) achieved better accuracy than uni-gram in both datasets, as shown in Table 17. Unlike gender classification, the designed features (All denotes all the features except uni-gram) are more effective for age-interval prediction. The prediction accuracy was enhanced by 12.55% on Taiwanese real names and 17.44% on FB names. This implies that additional features are especially important for age-interval prediction.

Table 17. Performance of the age-interval classifier on Taiwanese names.

Method	Real Name		FB Name	
	Year Error	Multi-Answer Acc.	Year Error	Multi-Answer Acc.
U	4.9711	0.5154	4.2200	0.4810
RFC_W	3.9785	0.6013	4.1331	0.4884
RFC_All	3.3878	0.7268	3.4264	0.6628

Multi-label classification can be regarded as one of the performance indicators, in which the model only needs to match one of the correct labels. Table 18 shows that the performance between the RFC and MLP is similar, but the performance of MLP dropped if more features were included.

Table 18. Multi-label results.

Features	Classifier	Recall	Precision	F1
All	RFC	0.637	0.731	0.681
WGF	MLP	0.627	0.745	0.681
All	MLP	0.611	0.742	0.670
U	MLP	0.598	0.744	0.660

4.3.2. Performance on Bilingual Names and Ethnic-Chinese Malaysian Names

Tables 19 and 20 below show that the features trained on Taiwanese names achieved slightly lower accuracy in gender prediction on Chinese names and age-interval prediction on bilingual names. Additionally, it follows a similar trend in which word embedding is more effective than uni-gram for gender and age-interval prediction. However, the performance of gender prediction on English names is poor. The possible reasons will be discussed in Section 5.2.

Table 19. Gender prediction on bilingual names.

Feature	Accuracy	
	Chinese Name	**English Name**
U	0.8511	0.2532
UP	0.7562	0.5182
W	0.9015	0.5651
WP	0.9039	0.5674
WUP	0.9050	0.5686

Table 20. Age-interval prediction on bilingual names.

	Feature	Average Year Error	
		Chinese Name	**English Name**
Unigram-based	U	5.2099	5.2170
	S	7.5407	7.5172
	M	6.8695	7.7489
	P	5.4941	5.7608
	US	5.1162	5.1243
	UM	5.0668	5.1058
	UP	4.9610	4.9939
Embedding-based	W	4.7100	4.7211
	WS	4.6881	4.6923
	WM	4.6889	4.6962
	WP	4.6872	4.6915
	WUP	4.6996	4.7051

There are two differences between ethnic-Chinese Malaysian names and standard English names. First, there are multiple Chinese dialects with different pronunciations used in Malaysia. Second, Malaysian names are usually disyllabic, written in two characters, but monosyllabic names, written in one character, are included as well. Therefore, the experiments were conducted based on the syllable: monosyllabic name denoted as *single name*; while disyllabic name denoted as *without single name.* The prediction results on Malaysian names were significantly worse than those on Taiwanese names and bilingual names, as shown in Table 21. Additionally, word-embedding methods performed poorly compared to unigram-based approaches in terms of average year errors, as presented in Table 21.

Table 21. Age-interval prediction on ethnic-Chinese Malaysian names.

	Feature	Average Year Error	
		Single Name	Without Single Name
Unigram-based	U	11.1470	10.8600
	S	11.1837	10.0149
	M	16.3608	15.0794
	P	14.7213	13.8360
	US	8.3997	8.3001
	UM	8.8870	7.2471
	UP	9.9148	9.4274
Embedding-based	W	17.9851	18.3118
	WP	14.7012	15.2240
	WUP	11.5443	11.5808

5. Discussion

5.1. Features' Impacts on Prediction Accuracy

The model we proposed is based on six features: word meaning (W), fortune-telling feature (F), Chinese zodiac (Z), word radical (R), word pronunciation (P), and gender probability (G). These features can be categorized into two groups: direct features and contextual features. The Direct Features include word meaning and word pronunciation, which come directly from name characters. Taking the Chinese word "美" as an example, the word meaning is beauty; the word pronunciation is "Mei". The Contextual Features include fortune-telling feature, Chinese zodiac, word radical, and gender probability, which originate from the Taiwanese cultural, social, and historic background. According to our experience, the Chinese word "美" is highly likely to belong to a female name.

The impact of these features on the prediction results can be discussed in three directions. First, considering how a single feature affects prediction results, the word meaning feature was found to be the most informative one. Using the word embedding feature alone, it can achieve 93.94% accuracy in gender prediction, as shown in Table 22, and an average error year of 3.9785 (equal to 0.80 age-interval) for age-interval prediction, as indicated in Table 23. However, the Chinese zodiac feature only provides 58.79% accuracy in gender prediction and generates an average error year of 8.8080 (equal to 1.76 age-interval) for age-interval prediction. The performance differences from different features are significant.

Table 22. Single feature for gender prediction.

Feature	Accuracy	F1 Score	Precision	Recall
Zodiac (Z)	0.5879	0.5918	0.9416	0.7268
Fortune-telling (F)	0.5955	0.5969	0.9504	0.7333
Radical (R)	0.8111	0.8256	0.8625	0.8437
Pronunciation (P)	0.8766	0.8844	0.9056	0.8949
Word embedding (W)	0.9394	0.9411	0.9571	0.9490

Table 23. Single feature for age-interval prediction.

Feature	Average Year Error	Multi-Answer Accuracy
Zodiac	8.8080	0.4277
Gender	8.5189	0.3729
Fortune-telling	8.1268	0.3617
Radical	6.8947	0.4394
Pronunciation	5.1444	0.5325
Word embedding	3.9785	0.6013

Second, if we utilize word-embedding with one-hot encoding as the basis, combining with only one other feature, the word pronunciation feature most successfully predicts gender, with 94.48% accuracy. This is very close to the highest level of 94.64% accuracy with multiple features. The combination of word embedding with other single features can achieve similar performance as well, as shown in Table 24. As for the age-interval prediction, the word-embedding feature combined with Chinese zodiac reaches the lowest average error year of 3.6481 (equal to 0.73 age-interval), as shown in Table 25. Therefore, word embedding with Chinese zodiac is the most effective feature pair for age estimation.

Table 24. Dual features for gender prediction.

Feature	Accuracy	F1 Score	Precision	Recall
WR	0.9391	0.9401	0.9578	0.9489
WF	0.9408	0.9436	0.9569	0.9502
WZ	0.9410	0.9439	0.9578	0.9508
WU	0.9433	0.9410	0.9640	0.9524
WP	0.9448	0.9460	0.9622	0.9540

Table 25. Dual features for age-interval prediction.

Feature	Average Year Error	Multi-Answer Accuracy
WR	3.9838	0.6003
WG	3.9774	0.5989
WP	3.9583	0.5995
WF	3.7356	0.6112
WZ	3.6481	0.7210

Finally, to achieve the best performance using all the available features, the feature combination of word embedding, Chinese zodiac, word radical, and unigram can reach the best accuracy—94.64% for gender prediction. However, interestingly, with only two features (word embedding and pronunciation), the model can also reach a similar accuracy of 94.48%. This also implies that, under the scenarios without Chinese contextual consideration, word pronunciation is quite informative for gender prediction, as shown in Table 26. As for the age-interval prediction, the minimum average year error of 3.3878 (equal to 0.68 age-interval) can be reached using all the features, as indicated in Table 27. Therefore, it can be concluded that, especially for complicated tasks such as age-interval prediction, contextual features effectively enhance the accuracy of prediction results besides the direct features. This finding further proves the importance of considering the contextual environment for prediction tasks when only minimal input is available.

Table 26. Gender prediction on Taiwanese names.

Feature	Accuracy	F1 Score	Precision	Recall
W	0.9394	0.9411	0.9571	0.9490
WF	0.9408	0.9436	0.9569	0.9502
WZ	0.941	0.9439	0.9578	0.9508
WPFZ	0.9413	0.9390	0.9628	0.9507
WPFZRU	0.9415	0.9412	0.9598	0.9504
WP	0.9448	0.9460	0.9622	0.9540
WPU	0.9458	0.9440	0.9664	0.9551
WZRU	0.9464	0.9442	0.9663	0.9551

Table 27. Age-interval prediction on Taiwanese names.

Feature	Average Year Error	Multi-Answer Accuracy
W	3.9785	0.6013
WZ	3.6481	0.7210
WF	3.7356	0.6112
WPFR	3.7293	0.6133
WZG	3.6610	0.7205
WFZ	3.5005	0.7308
WFZG	3.4971	0.7321
WPFZ	3.4480	0.7326
WPFZG	3.4452	0.7312
WPFZRG	3.3878	0.7268

5.2. The Model's Generalizability on Bilingual Names

To examine the generalizability of our model trained on Taiwanese names, we utilized the Taiwanese Public Figure Names dataset, which contains bilingual names—Chinese and English names for an individual. The difference between the two different languages is that features like fortune-telling, Chinese zodiac, or word radical cannot be applied to English names. To make a fair comparison, only the features useful for both languages were selected.

Table 19 demonstrates the performance comparison for gender prediction using different features. First, it can be observed that by using fewer features, the prediction accuracy on Chinese/bilingual names (90.5%) is lower compared to that on the Taiwanese Real Names dataset. One possible reason is that the names collected from Taiwanese public figures are more unique, which cannot represent the normal distribution of Taiwanese names. Additionally, as a public figure, the name may be changed for various reasons. Therefore, their names are less inclined to follow the naming rules described above.

Second, the gender prediction accuracy on English/bilingual names is significantly lower compared to the results on Chinese/bilingual names. This is because when a Taiwanese name is translated into English, some features may be missing. For example, both "偉宏"and "薇虹" are pronounced "Wei Hong" in English. However, they have different word meanings and radicals. The former one is usually used for a male name, while the latter is often for a female. Due to the missing features, the gender prediction accuracy on English names is just slightly better than a random guess. This implies that additional research is required to acquire more useful features.

Table 20 shows the performance comparison for age-interval prediction based on two different methods: unigram-based and embedding-based with one-hot encoding methods. To explore more features, we further separated the word pronunciation feature (P) into consonant feature (S) and vowel feature (M), as elaborated in the Feature Extraction section. In general, embedding-based approaches have less average year errors than unigram-based methods. Among all the features, word embedding combined with pronunciation achieved the minimum average year errors of 4.6872 (equal to 0.94 age-interval) and 4.6915 (equal to 0.94 age-interval) for Chinese names and English names, respectively. The prediction errors for bilingual names were slightly higher than those for Taiwanese names. Unlike the results of gender prediction, the differences in age-interval prediction between bilingual names were negligible. It can be regarded as the same level of performance. Therefore, word pronunciation (P) is an informative feature for age prediction on bilingual names.

One interesting question here is as follows: for a name translated from Chinese to English, why is word embedding combined with the pronunciation feature informative for age-interval estimation but not for gender classification? A possible explanation is the language preferences in different generations. In early generations, parents may have preferred using Taiwanese. More recently, parents mostly speak in standard Chinese. In other words, parents in different generations have different preferences in terms of choosing characters and their corresponding pronunciations. Therefore, the model can

still make useful age-interval prediction for English names after translation. However, the features useful for gender prediction were missing under the same scenario.

Therefore, it can be concluded that the generalizability of our model can be successfully extended to bilingual language for age-interval prediction. However, for gender prediction, we need additional features for English names.

5.3. The Model's Cross-Border Generalizability on Ethnic-Chinese Malaysian Names

The model's cross-border learning generalizability was further examined on a dataset collected from Malaysia: ethnic-Chinese Malaysian names, expressed in English. From Table 21, it can be easily observed that all the average year errors from various single features or feature combinations are significantly larger than the results from previous experiments. The best performance of average year error of 7.2471 (equal to 1.45 age-interval) using uni-gram combined with the vowel feature is 2.14 times larger (average year error of 3.3878) than that on Taiwanese names. The worst performance can even reach an average year error of 18.3118 (equal to 3.66 age-interval).

The main reason is that multiple Chinese dialects used in Malaysia have different pronunciations. For instance, the surname "陳" is pronounced "Chen" in Chinese, but "Tan" in Hokkien, one of the Chinese dialects, as shown in Table 28. Although the previous experiments on bilingual names received satisfactory results concerning age-interval prediction for Chinese and English names, the model did not present similar generalizability for the English names collected in Malaysia. In other words, the word pronunciation feature extracted from Taiwanese names using standard Chinese does not fit into Malaysian names pronounced in other Chinese dialects. Considering the large performance gap with the previous results, it is claimed that, unfortunately, the model trained on Taiwanese names cannot be directly applied to ethnic-Chinese Malaysian names. Thus, additional research exploring other features based on Malaysian names should be conducted in the future.

Table 28. Different pronunciation in different dialects.

Dialect	Pronunciation
Chinese	Chen
Cantonese	Chan
Hokkien	Tan
Teochew	Tan

6. Conclusions

In this paper, making gender and age interval predictions from minimal data by considering the contextual environment of a given name was demonstrated. It showed that not only data itself but also its contextual factors could be the input to generate valuable outcomes. The designed features—such as Chinese zodiac, fortune-telling feature, and word radical—provide insights into how these factors improve prediction accuracy based on cultural, social, and historical understandings.

In contrast to previous studies, we presented a complete framework utilizing direct information and contextual consideration as input for making predictions. This framework provides insights into how AI can create value under the scenarios of limited available data. Furthermore, the generalizability of the model was examined using bilingual and cross-border datasets. The results show satisfactory prediction accuracy for Chinese names. However, the different pronunciations of multiple Chinese dialects used in Malaysia lowered the transferability of the model trained on standard Chinese. This implies that further studies exploring the additional features of different languages could be a worthwhile future research direction.

Author Contributions: Conceptualization, J.-S.L., Y.-S.C. and Y.-H.H.; methodology, C.-Y.H. and Y.-H.H.; software, C.-Y.H. and L.-Y.C.; validation, C.-Y.H. and L.-Y.C.; formal analysis, J.-S.L.; investigation, J.-S.L.; data curation, C.-Y.H. and L.-Y.C.; writing—original draft preparation, J.-S.L.;

writing—review and editing, J.-S.L. and Y.-S.C.; visualization, J.-S.L.; supervision, Y.-S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology in Taiwan, grant number "MOST 110-2221-E-007-085-MY3" and "MOST 108-2221-E-007-064-MY3". The funding is to facilitate general study in the domain of Artificial Intelligence, not for specific topic.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this research are involved with personal privacy. They are utilized for research purposes only, not shared in the public domain.

Acknowledgments: Thanks for Chih-Hao Tsai, currently the chairman of UiGathering, for providing "A List of Chinese Names" as part of Taiwanese names.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805v2.
2. Zhou, C.; Sun, C.; Liu, Z.; Lau, F. A C-LSTM Neural Network for Text Classification. *arXiv* **2015**, arXiv:1511.08630.
3. General Data Protection Regulation (GDPR). 2016, ver. OJ L 119, 04.05, European Commission Website. Available online: https://gdpr-info.eu/ (accessed on 22 June 2021).
4. A Close Reading of China's Data Security Law, in Effect Sept. 1, 2021. China Briefing. 2021. Available online: https://www.china-briefing.com/news/a-close-reading-of-chinas-data-security-law-in-effect-sept-1-2 (accessed on 3 August 2021).
5. China Passes New Personal Data Privacy Law, to Take Effect Nov. 1. Reuters. 2021. Available online: https://www.reuters.com/world/china/china-passes-new-personal-data-privacy-law-take-effect-nov-1-2021-08-20/ (accessed on 8 September 2021).
6. Koene, A. Algorithmic Bias: Addressing Growing Concerns. *IEEE Technol. Soc. Mag.* **2017**, *36*, 31–32. [CrossRef]
7. Courtland, R. The Bias Detectives. *Nature* **2018**, *558*, 357–360. [CrossRef]
8. Wilson, H.J.; Daugherty, P.R.; Davenport, C. Harvard Business Review. The Future of AI Will Be about Less Data, Not More. 2019. Available online: https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more (accessed on 9 August 2021).
9. Barry, H., Jr.; Harper, A.S. Increased choice of female phonetic attributes in first names. *Sex Roles* **1995**, *32*, 809–819. [CrossRef]
10. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781v3.
11. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
12. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365.
13. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Cassification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 1480–1489.
14. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. Truth of Varying Shades: Analyzing language in Fake News and Political Factchecking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 2931–2937.
15. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
16. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems, Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, QC, Canada, 8–14 December 2019*; Curran Associates: Red Hook, NY, USA, 2020; pp. 5754–5764.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates: Red Hook, NY, USA, 2018; pp. 5998–6008.
18. Knowles, R.; Carroll, J.; Dredze, M. Demographer: Extremely Simple Name Demographics. In Proceedings of the First Workshop on NLP and Computational Social Science, Austin, TX, USA, 5 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 108–113.
19. Mueller, J.; Stumme, G. Gender Inference Using Statistical Name Characteristics in Twitter. In Proceedings of the 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016, Union, NJ, USA, 15–17 August 2016; Data Science: Union, NJ, USA, 2016; pp. 1–8.

20. Burger, J.D.; Henderson, J.; Kim, G.; Zarrella, G. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–29 July 2011*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 1301–1309.
21. Brown, E. Gender Inference from Character Sequences in Multinational First Names. 2017. Available online: https://towardsdatascience.com/name2gender-introduction-626d89378fb0#408a (accessed on 18 August 2021).
22. Hu, Y.; Hu, C.; Tran, T.; Kasturi, T.; Joseph, E.; Gillingham, M. *What's in a Name?- Gender Classification of Names with Character Based Machine Learning Models*; ACM: New York, NY, USA, 2021.
23. Malmasi, S.; Dras, M. A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations. In Proceedings of the Australasian Language Technology Association Workshop, Melbourne, VIC, Australia, 26–28 November 2014; Australasian Language Technology Association Workshop (ALTA): Melbourne, VIC, Australia, 2014; pp. 145–149.
24. Wood-Doughty, Z.; Andrews, N.; Marvin, R.; Dredze, M. Predicting Twitter User Demographics from Names Alone. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 105–111.
25. Wong, K.O.; Zaïane, O.R.; Davis, F.G.; Yasui, Y. A machine learning approach to predict ethnicity using personal name and census location in Canada. *PLoS ONE* **2020**, *15*, e0241239. [CrossRef]
26. Lee, J.; Kim, H.; Ko, M.; Choi, D.; Choi, J.; Kang, J. Name Nationality Classification with Recurrent Neural Networks. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, VIC, Australia, 19–25 August 2017.
27. Nguyen, D.; Smith, N.A.; Rose, C. Author Age Prediction from Text Using Linear Regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011; Association for Computational Linguistics: Portland, OR, USA, 2011; pp. 115–123.
28. Abdallah, E.E.; Alzghoul, J.R.; Alzghool, M. Age and Gender prediction in Open Domain Text. *Procedia Comput. Sci.* **2020**, *170*, 563–570. [CrossRef]
29. Peersman, C.; Daelemans, W.; Van Vaerenbergh, L. Predicting Age and Gender in Online Social Networks. In Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, SMUC '11, Glasgow, UK, 28 October 2011; ACM: New York, NY, USA, 2011; pp. 37–44.
30. Rosenthal, S.; McKeown, K. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-social Media Generations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1, HLT '11, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 763–772.
31. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Booth. Linguistic Inquiry and Word Count: Liwc 2001. *Mahway Lawrence Erlbaum Assoc.* **2001**, *71*, 2001.
32. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.; Dziurzynski, L.; Ramones, S.M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M.E.P.; et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* **2013**, *8*, e73791. [CrossRef] [PubMed]
33. Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Stillwell, D.; Kosinski, M.; Ungar, L.; Schwartz, H.A. Developing Age and Gender Predictive Lexica over Social Media. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1146–1151.
34. Bukar, A.M.; Ugail, H.; Connah, D. Automatic age and gender classification using supervised appearance model. *J. Electron. Imaging* **2016**, *25*, 061605. [CrossRef]
35. Ferguson, E.; Wilkinson, C. Juvenile age estimation from facial images. *Sci. Justice* **2017**, *57*, 58–62. [CrossRef]
36. Fu, Y.; Guo, G.; Huang, T.S. Age Synthesis and Estimation via Faces: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976.
37. Geng, X.; Yin, C.; Zhou, Z.-H. Facial Age Estimation by Learning from Label Distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2401–2412. [CrossRef] [PubMed]
38. Chao, W.-L.; Liu, J.-Z.; Ding, J.-J. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognit.* **2013**, *46*, 628–641. [CrossRef]
39. Pontes, J.K.; Britto, A.S.; Fookes, C.; Koerich, A.L. A flexible hierarchical approach for facial age estimation based on multiple features. *Pattern Recognit.* **2016**, *54*, 34–51. [CrossRef]
40. Choi, S.E.; Lee, Y.J.; Lee, S.J.; Park, K.R.; Kim, J. Age Estimation using a Hierarchical Classifier based on Global and Local Facial Features. *Pattern Recognit.* **2011**, *44*, 1262–1281. [CrossRef]
41. Huerta, I.; Fernández, C.; Prati, A. *Facial Age Estimation through the Fusion of Texture and Local Appearance Descriptors. ECCV 2014 Workshops*; Part II, LNCS 8926; Springer International Publishing Switzerland: Basel, Switzerland, 2015; pp. 667–681.
42. Abousaleh, F.S.; Lim, T.; Cheng, W.-H.; Yu, N.-H.; Hossain, M.A.; Alhamid, M.F. A novel comparative deep learning framework for facial age estimation. *EURASIP J. Image Video Process.* **2016**, *2016*, 47. [CrossRef]
43. Hsiao, C.Y.; Huang, Y.H.; Calderon, F.H.; Chen, Y.S. Behind the Name: A Comparative Framework for Age Estimation of Taiwanese Names. In Proceedings of the 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taipei, Taiwan, 3–5 December 2020.
44. Tsai, C.-H. A List of Chinese Names. Available online: http://technology.chtsai.org/namelist/ (accessed on 20 January 2020).

45. Melton, J.G. *The Encyclopedia of Religious Phenomena*; Visible Ink Press: Canton, MI, USA, 2008.
46. Chen, G.Y. *Feng-Shui Bible-Lucky Farmer's Calendar*; Caishe International Co., Ltd.: Taipei, Taiwan, 2018.
47. Yang, Z.Y. *The Best Naming Study*; Chilin Culture Publishing House: Taipei, Taiwan, 2010.
48. Grech, V. The Influence of the Chinese Zodiac on the Male-to-female Ratio at Birth in Hongkong. *J. Chin. Med. Assoc.* **2015**, *78*, 287–291. [CrossRef] [PubMed]
49. Kiel/Lsuni International Phonetic Alphabet. (Revised to 2020). Available online: https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf (accessed on 17 August 2021).