Expert Systems With Applications 64 (2016) 423–433

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An on-device gender prediction method for mobile users using representative wordsets

Yerim Choi*, Yoonjung Kim, Solee Kim, Kyuyon Park, Jonghun Park

Department of Industrial Engineering, Seoul National University, Seoul, 151-744, Korea

ARTICLE INFO

Article history:
Received 17 April 2016
Revised 21 July 2016
Accepted 1 August 2016
Available online 3 August 2016

Keywords:
Gender prediction
Mobile text data
Representative wordsets
Word evaluation measures
On-device analytics

ABSTRACT

With the proliferation of mobile devices and the growing necessity for gender information in personalized intelligent systems, gender prediction of mobile users has become an important research issue. Text data in mobile devices are known to have high discriminative power for gender, but transmitting those data to the outside of a device has a security risk and raises a privacy concern of users. This study introduces an on-device gender prediction framework, by which the entire data analysis is performed inside a device minimizing the privacy risk. To cope with the resource limitation of mobile devices, gender information of a user is predicted by matching the user's mobile text data against gender representative wordsets which are constructed from web documents using a word evaluation measure. From the experiments conducted on real-world datasets, the effectiveness of the proposed framework was confirmed, and it was concluded that not only discriminability of a word but also popularity should be considered for the on-device gender prediction. The proposed framework is simple yet very powerful for gender prediction that its practical application to various expert and intelligent systems is possible attributed to the low computational complexity and high prediction performances.

1. Introduction

Recently, gender prediction research for mobile device users has been widely conducted due to the growing importance of demographic information of users in personalized systems. For instance, Choi, Oh, Kim, and Ryu (2016) reported that the accuracy of recommendations differs depending on the gender of a user, and gender was considered as one of the key factors in diverse studies (Chong, 2013; Yeh, Wang, & Yieh, 2016). Gender prediction of a mobile device user is a problem of inferring the user's gender by analyzing her/his mobile usage data including location, sensor, and text data. Solving the problem has become possible as a large amount of usage data are collected from mobile devices.

Text data are proven to be effective for gender prediction from the previous studies utilizing user-generated text data in instant messaging (Huang, Li, & Lin, 2014) and social network services (SNSs) (Alowibdi, Buy, & Yu, 2013b; Ikeda, Hattori, Ono, Asoh, & Higashino, 2013; López-Monroy, Montes-y Gómez, Escalante, Villaseñor-Pineda, & Stamatatos, 2015). The effectiveness is attributed to the implicit and explicit differences in the text female and male uses. Nunes, Miles, Luck, Barbosa, and Lucena (2015) noted that the text data are useful for the gender prediction as the preferences of a user are reflected in the words he or she uses, and the difference in interests of females and males may implicitly affect the selection of words they use. Moreover, in some languages, females and males explicitly use different words. For instance, in Korean, there are different appellations for a younger female calling an older male and a younger male calling an older female, and in Latin languages, different word endings are used by females and males.

In this regard, it also has been reported that there exist significant differences in the mobile text data generated by females and males (Igarashi, Takai, & Yoshida, 2005). The mobile text includes text messages, contacts, web document bookmarks, and search keywords. Although utilizing the mobile text data would results in accurate gender classification, collecting the data is often unacceptable to many users due to the privacy concern and usually illegal as the data contain highly private information. One viable solution for addressing the privacy issue is to perform data analysis on-device, meaning that the analysis is carried out entirely within a mobile device. Mukherji, Srinivasan, and Welbourne (2014) utilized the on-device data analysis in an effort to reduce the risk of transferring location information to outside.

However, it is not practical to perform text data processing on-device since the text processing is infamous for its heavy

* Corresponding author.
E-mail addresses: iangoozh@gmail.com (Y. Choi), yoonj625@gmail.com (Y. Kim), kpsinw@gmail.com (S. Kim), mysnuky91@snu.ac.kr (K. Park), jonghun@snu.ac.kr (J. Park).

http://dx.doi.org/10.1016/j.eswa.2016.08.002
0957-4174/

Table 1
Summary of previous research on gender prediction.

Feature type	Domain	Feature	Publication
Textual	Web service	User names	Alowibdi et al. (2013a)
		Posts and comments in SNSs	López-Monroy et al. (2015)
			Huang et al. (2014)
			Ikeda et al. (2013)
			Miller et al. (2012)
		Email contents	Deitrick et al. (2012)
			Cheng et al. (2011)
		Chat messages	Kucukyilmaz et al. (2006)
	Mobile device	Installed applications	Seneviratne et al. (2014)
Non textual	Web service	SNS profile page colors	Alowibdi et al. (2013b)
		Web browsing patterns	Hajmohammadi et al. (2014)
		Social networks	Peersman et al. (2011)
	Mobile device	Sensor data	Zhong et al. (2013)
			Ying et al. (2012)
		Social networks	Dong et al. (2014)
		Accelerometer data	Weiss and Lockhart (2011)

computation due to the high dimensionality of text data representation (Sebastiani, (2002)) while the storage size and computing power of mobile devices are limited (Barragáns-Martínez, Costa-Montenegro, & Juncal-Martínez, 2015). Although it is not impossible to embed such advanced methods in mobile devices, there exists trade-off between accuracy and resource usage (Han et al., 2016). To address the resource limitation, Shotton et al. (2013) proposed decision jungle which is compact version of random forest to embed classification models inside small devices. In addition, most of the previous on-device analytics studies were focused on reducing the complexity of existing methods (Jeong, Cheng, Song, & Kalasapur, 2009) or implementing low complexity methods such as frequent rule mining (Srinivasan et al., 2014) and Naïve Bayes (Mukherji et al., 2014)

In an attempt to address the computational concerns of text data processing, research on word selection has been actively pursued. Instead of using all unique words for text data representation, words are evaluated according to purposes and selected based on the evaluation results. For instance, class discriminability of words is evaluated for the text classification (Rogati & Yang, 2002), while topic specificity of words is measured for the keyword extraction (Noh, Jo, & Lee, 2015).

Motivated by the above remarks, we propose a novel gender prediction framework which performs on-device data analysis by matching mobile text data against gender representative wordsets, each of which is composed of words representing a gender, female or male. The representativeness of a word is evaluated by several word evaluation measures, and the considered characteristics of representativeness differ depending on the measures. The proposed framework is composed of two phases, representative wordset construction phase, where the most representative words for each gender are selected, and on-device gender prediction phase, where the gender of a user is predicted by comparing similarities between the user's mobile text data and the representative wordsets.

Specifically, we have conducted a comparative study on the word evaluation measures to determine which characteristics of a word are effective for on-device gender prediction. Among various word evaluation measures, five measures which have been widely used in previous literature, namely DF (document frequency), TF-IDF (term frequency-inverse document frequency), MI (mutual information), CHI (chi-square), and BNS (bi-normal separation), and four newly utilized measures are compared. In addition, four kinds of similarity weighting schemes, which are the combinations of two word matching types, binary matching and frequency counting, and two word scoring types, binary scoring and word evaluation value, are employed for the performance comparisons.

The paper is organized as follows. Previous gender prediction methods and word evaluation measures are introduced in Section 2. In Section 3, the proposed framework is explained in detail, and results of experiments conducted to examine the performances of the proposed framework using various word evaluation measures are presented in Section 4. Lastly, the paper is concluded in Section 5.

2. Literature review

2.1. Gender prediction

Previous gender prediction studies can be categorized into two groups according to the data they utilized, text-based gender prediction and non-text-based gender prediction, as shown in Table 1. Traditionally, mostly in web services, user generated texts such as user names (Alowibdi, Buy, & Yu, 2013a), posts and comments in SNSs (Huang et al., 2014; Ikeda et al., 2013; López-Monroy et al., 2015; Miller, Dickinson, & Hu, 2012), e-mail contents (Cheng, Chandramouli, & Subbalakshmi, 2011; Deitrick et al., 2012), and chat messages (Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2006) were utilized for gender classification, and have shown a successful performance attributed the inherent gender discriminability of text data. They utilize a number of discriminative words to maximize the gender prediction performance. Kucukyilmaz et al. (2006) employed 3,000 words from chat mining for the prediction, and Miller et al. (2012) concluded that utilizing 15,000 n-gram features from Twitter messages worked the best for the prediction. Moreover, heavy methods were adopted to extend the coverage of the selected words, such as query extension, ontology based approaches, and semantic word search, which were too heavy to be embedded in mobile devices. The proposed method is very light since it only utilizes a small number of representative words for classification, but if the words are not founded in the texts generated by a user, the gender of the user cannot be determined.

Unlike gender prediction for web service users, no textual feature has been used in gender prediction for mobile devices users except for package names of installed applications (Seneviratne, Seneviratne, Mohapatra, & Mahanti, 2014). Non-textual features utilized in web services include colors of SNS profile page (Huang et al., 2014), web browsing patterns (Hajmohammadi, Ibrahim, & Selamat, 2014), and social networks (Peersman, Daelemans, & Van Vaerenbergh, 2011). Although the non-textual features have an advantage that they can be used independently of languages, predictions using textual features showed better performances with affordable complexities compared to the methods using non-textual features (Huang et al., 2014).

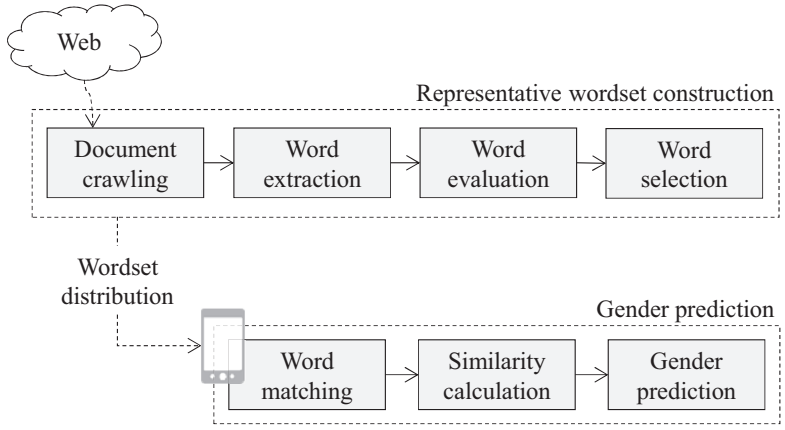


Fig. 1. Overview of the proposed on-device gender prediction method using representative wordsets.

Table 2
Summary of previous research utilizing word evaluation measures.

Purpose	Word evaluation measure	Publication
Document search	TF-IDF, MI	Roul et al. (2014) Loscalzo et al. (2014)
Keyword extraction	TF, IDF, TF-IDF	Noh et al. (2015) Litvak and Last (2008) Azcarraga et al. (2012)
Text classification	MI, CHI, BNS	Forman (2003) Rogati and Yang (2002) Lu et al. (2010) Yang and Pedersen (1997)
Email classification	TF-IDF, CHI	Wang et al. (2015)
Gender prediction	MI, CHI	Miller et al. (2012)

Recently, research were performed to detection the gender of a mobile device user using data inside the device such as sensor data (Ying, Chang, Huang, & Tseng, 2012; Zhong, Tan, Mo, & Yang, 2013), social networks constructed from contact information (Dong, Yang, Tang, Yang, & Chawla, 2014), and accelerometer data (Weiss & Lockhart, 2011). Due to the privacy concern, only less private data were allowed to be transferred to a server to perform the gender prediction. The gender prediction performances using those data were low compared to the traditional text-based approaches. By enabling the utilization of text data inside a mobile device, the proposed framework predicts the gender of a user with higher accuracy, while requiring low computing power and memory resource.

2.2. Word evaluation measures

In natural language processing domain, word evaluation measures have been developed for diverse purposes including web document search (Loscalzo, Wright, & Yu, 2014; Roul, Devanand, & Sahay, 2014), keyword extraction (Azcarraga, Liu, & Setiono, 2012; Litvak & Last, 2008; Noh et al., 2015), and word selection for document classification (Chung, 2014; Forman, 2003; Lu, Chiang, Keh, & Huang, 2010), email classification (Wang, Liu, Feng, & Zhu, 2015) or gender prediction (Miller et al., 2012). In these studies, each measure reflects different characteristics of a word depending on the evaluation purpose. For instance, specificity of a word to a given topic was a major criterion for keyword extraction (Roul et al., 2014), while discriminative power was a major criterion for classification (Forman, 2003). Table 2 shows a list of previous studies with their word evaluation purposes and employed measures.

For a web document retrieval problem, topic specificity of a word has been evaluated by using TF-IDF (Roul et al., 2014) and MI (Loscalzo et al., 2014). TF-IDF was also used for keyword extraction problem (Noh et al., 2015), along with TF and DF (Azcarraga et al., 2012), which respectively evaluate the number of a word's occurrences in the whole set of documents and the number of documents where the word appears. For a classification problem, discriminative power of a word against a class was evaluated, and well known measures include CHI (Rogati & Yang, 2002; Wang et al., 2015), MI (Yang & Pedersen, 1997), and BNS (Forman, 2003). Particularly, for gender prediction of a user, CHI and MI were adopted (Miller et al., 2012). In this study, we compare diverse word evaluation measures to find appropriate characteristics of representative words for on-device gender prediction without compromising the privacy concerns.

3. On-device gender prediction

3.1. Problem definition

In this paper, we attempt to predict the gender of a mobile device user by comparing mobile text data of the user with representative wordsets derived from web documents. Mobile text data of user, u, is represented as a character sequence, denoted by T_u, which is an aggregation of texts in the user's device such as text messages, search keywords, and web document bookmarks. T_u is represented as a long character sequence in order to avoid the time consuming word extraction procedure.

Gender is denoted by g, and it has one of the two constants, `female` and `male`. The set of web documents written by authors of g is denoted by $\mathcal{D}_g$, and its element, D_g, is a document written by an author of g. In contrast to T_u, D_g is represented as a bag of words, composed of the words parsed from a raw web document. Each word is denoted as w. We let N_g be the total number of D_g in $\mathcal{D}_g$, and N be the summation of N_g for all $g \in \{\texttt{female}, \texttt{male}\}$.

By analyzing $\mathcal{D}_g$, a representative wordset for g is constructed, and it is denoted as W_g. W_g is composed of representative words which are selected according to a word evaluation result. Subsequently, word evaluation is performed to all the unique words which appear in $\mathcal{D}_g$ for all g's.

3.2. Overview

The proposed method is composed of two phases, representative wordset construction and on-device gender prediction phases as shown in Fig. 1. In the representative wordset construction phase, four steps are performed. First, web documents whose authors' genders are known are collected, and they are represented as a bag of words through word extraction process (Litvak & Last, 2008). Then, gender representativeness of a word is calculated by using a word evaluation measure, and most representative

words for g are selected according to the evaluation score. Finally, the selected words form W_g.

Inside a device, three steps of the on-device gender prediction phase are performed using W_g for $g \in \{\texttt{female}, \texttt{male}\}$. First, w in W_g is matched to T_u to check the existence of w in T_u. Then, similarities between the matching result and W_g for all g's are calculated. The gender of the mobile device user, u, is predicted as g with higher similarity.

3.3. Representative wordset construction

3.3.1. Document crawling and word extraction

For representative wordset construction, we collect web documents whose authors' gender is known. While gender information is not available in most web documents, some services provide such information. For instance, social network and blog services let users present their gender information in profile pages, and there are web community services which are open to only the users of a certain gender. Words are extracted from the crawled documents by tokenizing a document and converting them into root form using dictionaries (Noh et al., 2015). Lastly, D_g is represented as a bag of the extracted words.

3.3.2. Word evaluation

To select representative words, each unique word, w, in $\mathcal{D}_g$ is evaluated by using several word evaluation measures. They include some of well-known measures such as DF, TF-IDF, MI, CHI, and BNS and newly utilized measures, CE and AE. Each measure considers different aspects of the representativeness, namely popularity, specificity, and discriminability, respectively indicating frequency of w in D_g, degree of relevance of w to a specific topic, and whether w has a power of distinguishing g or not.

DF and TF-IDF. DF is a measure for counting the number of D_g containing w, and it represents the popularity of w. For g and w, DF_{gw} is defined by Eq. (1).

$$DF_{gw} = \sum_{D_g \in \mathcal{D}_g} E(D_g, w), \tag{1}$$

where $E(D_g, w)$ is a function indicating whether w exists in D_g as in Eq. (2).

$$E(D_g, w) = \begin{cases} 1, & \text{if } w \text{ exists in } D_g, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

TF-IDF is a ratio of the number of occurrences of w in $\mathcal{D}_g$ to the number of D_g containing w. The high TF-IDF score of w implies that w is topic specific (Aizawa, 2003). Eq. (3) defines TF-IDF score for g and w.

$$TF\text{-}IDF_{gw} = \frac{TF_{gw}}{DF_{gw}}, \tag{3}$$

where TF_{gw} is the number of occurrences of w across $\mathcal{D}_g$ as in Eq. (4).

$$TF_{gw} = \sum_{D_g \in \mathcal{D}_g} C(D_g, w), \tag{4}$$

where $C(D_g, w)$ is a function which returns the number of appearances of w in D_g. Note that, in this paper, normalized values of TF and DF, whose summations across all w are one, are used to address the scale imbalance problem of N_g for $g \in \{\texttt{female}, \texttt{male}\}$.

MI. MI is a measure of the mutual dependency between w and g. In detail, MI employs entropy to measure how much w is related to g as shown in Eq. (5).

$$MI_{gw} = \frac{N \times DF_{gw}}{(DF_{gw} + DF_{\bar{g}w})(DF_g + DF_{g\bar{w}})}, \tag{5}$$

where $\bar{g}$ indicates female when g is male, and vice versa. Similarly, $DF_{(g\bar{w})}$ is the number of D_{gw} not containing w, as defined by Eq. (6). Note that, we obtain values of TF-IDF and MI without logarithm transformation as the results with and without the transformation are rank equivalent and we only are interested in the ranks of words.

$$DF_{g\bar{w}} = N_g - DF_{gw}. \tag{6}$$

When w is independent of g, MI_{gw} is zero.

CHI. CHI measures the dependency between w and g, which is reflected in the difference between the observed number of occurrences and the expected number of occurrences of w when the occurrences are assumed to be independent of g. A CHI score for g and w is calculated by Eq. (7).

$$CHI_{gw} = \begin{cases} \frac{N \times \delta^2}{(DF_{gw}+DF_{g\bar{w}})(DF_{\bar{g}w}+DF_{\bar{g}\bar{w}})(DF_{gw}+DF_{\bar{g}w})(DF_{g\bar{w}}+DF_{\bar{g}\bar{w}}),} & \text{if } \delta > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\delta = DF_{gw}DF_{\bar{g}\bar{w}} - DF_{\bar{g}w}DF_{g\bar{w}}$. When w is independent of g, CHI_{gw} is zero. CHI has a drawback that it becomes unreliable when the expected number of occurrences is too small.

BNS. BNS is a word evaluation measure proposed by Forman (2003). It reflects the discriminability of w and is calculated by Eq. (8).

$$BNS_{gw} = \begin{cases} \left| F^{-1}\left(\frac{DF_{gw}}{DF_{gw}+D_{g\bar{w}}}\right) - F^{-1}\left(\frac{DF_{\bar{g}w}}{DF_{\bar{g}w}+DF_{\bar{g}\bar{w}}}\right) \right|, & \text{if } \delta > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where F^{-1} is an inverse cumulative probability function for a standard normal distribution. The number of occurrences of w is assumed to be normally distributed, and the prevalence rates for two genders are compared. The details are presented in Forman (2003).

CE and AE. In addition to the word evaluation measures mentioned above, which were successfully employed in previous gender prediction research for web documents, we utilize two word evaluation measures, CE and AE, to explore the characteristics necessary for representative words for on-device gender prediction using T_u. CE considers words that only appear in documents by one gender, while AE permits words used by one gender to occur in the documents by the other gender up to certain times. CE and AE take DF or TF-IDF of w as an evaluation score. In this regard, the effects of popularity and topic specificity of w for on-device gender prediction can be evaluated while discriminability is considered.

CE considers w which only appears in documents of one gender, and the value of CE is determined as the value of DF or TF-IDF of w according to its type, denoted as DF(CE) or TF-IDF(CE). Given g and w, DF(CE) and TF-IDF(CE) scores are calculated by Eqs. (9) and (10), respectively.

$$DF(CE)_{gw} = \begin{cases} DF_{gw}, & \text{if } DF_{\bar{g}w} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

$$TF\text{-}IDF(CE)_{gw} = \begin{cases} TF\text{-}IDF_{gw}, & \text{if } DF_{\bar{g}w} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

In the same manner, two types of AE are utilized, DF(AE) and TF-IDF(AE). Unlike CE, AE allows w to occur in both female and male documents for the limited number of times. DF(AE) and TF-IDF(AE) scores are calculated by Eqs. (11) and (12), respectively.

$$DF(AE)_{gw} = \begin{cases} DF_{gw}, & \text{if } \frac{DF_{\bar{g}w}}{N_{\bar{g}}} < \rho, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

$$TF\text{-}IDF(AE)_{gw} = \begin{cases} TF\text{-}IDF_{gw}, & \text{if } \frac{DF_{\bar{g}w}}{N_{\bar{g}}} < \rho, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where ρ is the predefined value which is the ratio of allowed $DF_{\bar{g}w}$ to $N_{\bar{g}}$.

Table 3
Summary of the collected web documents and mobile text data.

Data type	Category	Total	Female	Male
Web documents	The number of documents	189,127	53,382	135,745
	The number of unique words	141,509	117,941	137,743
	The number of words used by only one gender	–	3,766	23,568
	The average document length (in word)	317	359	301
Mobile text data	The number of users	32	16	16
	The average text data length (in character)	89,188	124,945	53,432

3.3.3. Word selection

From the word evaluation results, words are selected for representative wordset construction. W_g is composed of R representative words, and each representative word in W_g is denoted as w. Specifically, top R words from unique words in $\mathcal{D}_g$ for all $g \in \{\texttt{female}, \texttt{male}\}$ are selected according to the word evaluation scores, $s(w)$, where R is a predefined number, same for both W_g's.

3.4. On-device gender prediction using representative wordsets

3.4.1. Word matching

Inside a device of u, gender prediction is performed using W_g through a word matching procedure. During the procedure, T_u and w are matched to check the existence of w in T_u as in Eq. (13).

$$E(T_u, w) = \begin{cases} 1, & \text{if } w \text{ exists in } T_u \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where $E(T_u, w)$ is a function which returns 1 when w is found in T_u, and 0, otherwise. We call this type of word matching scheme as binary matching. The matching result is represented as an R-dimensional binary vector after repeating the procedure for all $w \in W_g$, and it is denoted by $\vec{e}_{gu}$, whose element indicates whether w exists in T_u as in Eq. (14).

$$\vec{e}_{gu} = \langle E(T_u, w) \rangle, \text{ where } \vec{e}_{gu} \in \{0, 1\}^R. \tag{14}$$

In addition to the binary matching, we consider frequency counting, in which the occurrences of w in T_u is utilized. As a result, an R-dimensional integer vector, $\vec{f}_{gu} \in \mathbb{I}^R$, whose element is the matching frequencies between w and T_u, is defined by Eq. (15).

$$\vec{f}_{gu} = \langle C(T_u, w) \rangle, \tag{15}$$

where $C(T_u, w)$ is a function which counts the number of matches between w and T_u.

Note that the same symbol E (and C, respectively) is used in Eqs. (1) and (14) (and Eqs. (4) and (15), respectively) for the sake of simplicity although different argument types are supplied. While both T_u and D_g represent text data, T_u is a character sequence, and D_g is a bag of words, which is acquired by performing word extraction process to a character sequence. Therefore, a function that takes T_u as an input requires less computation than that takes D_g since word extraction process is skipped.

3.4.2. Similarity calculation

After the word matching step, similarities between T_u and W_g for $g \in \{\texttt{female}, \texttt{male}\}$ are calculated by using cosine similarity, $sim(W_g, T_u)$ which is defined as Eq. (16).

$$sim(W_g, T_u) = \frac{\vec{v}_g \cdot \vec{v}_{gu}}{|\vec{v}_g||\vec{v}_{gu}|}, \tag{16}$$

where $\vec{v}_g$ and $\vec{v}_{gu}$ are vector representations of W_g and T_u for g, respectively, and $\vec{v}_g \cdot \vec{v}_{gu}$ is the inner product between them.

Specifically, $\vec{v}_g$ is a vector representing scores of words of W_g. We consider two types of scores: binary score which is represented as an R-dimensional one vector (denoted as $\vec{1}$) indicating the existences of all words in W_g, and word evaluation score (denoted by $\vec{s}_g$) defined in Eq. (17).

$$\vec{s}_g = \langle s(w) \rangle, \text{ where } w \in W_g \text{ and } s_g \in \mathbb{R}_+^R. \tag{17}$$

$s(w)$ is the word evaluation score for w such as MI and CHI introduced in Section 3.3.2. On the other hand, $\vec{v}_{gu}$ is a vector representing word matching results for given W_g and T_u, and is either $\vec{e}_{gu}$ in Eq. (14) or $\vec{f}_{gu}$ in Eq. (15) depending on the word matching scheme. As a result, there are four combinations of similarities according to the vector representations of W_g and T_u, leading to binary score and binary matching (BB), binary score and frequency counting (BF), word evaluation score and binary matching (EB), and word evaluation score and frequency counting (EF). Note that the elements of the four vectors are normalized to make their sum across all $w \in W_g$ become one.

3.4.3. Gender prediction

Gender of a user is predicted as a gender with higher similarity among $sim(W_g, T_u)$ for $g \in \{\texttt{female}, \texttt{male}\}$. The predicted gender of the user, $\hat{g}$, is determined by Eq. (18).

$$\hat{g} = \text{argmax}_g sim(W_g, T_u), \ \forall g \in \{\texttt{female}, \texttt{male}\}, \tag{18}$$

where $sim(W_g, T_u)$ is calculated by using Eq. (16).

4. Experiment

4.1. Datasets

We conducted experiments using real-world datasets. Table 3 shows a summary of the datasets. Blog posts were collected as the web documents for representative wordset construction, and mobile text data were acquired for the prediction performance evaluation. Total of 189,127 documents were collected from 162 Korean blogs, where the genders of authors are known. Mobile text data were collected from 32 Android mobile device users composed of the equal number of females and males by using an Android application developed to collect text messages, web site bookmarks, and search keywords.

Words extracted by both Korean text analyzers, Kind Korean Morpheme Analyzer (KKMA) (Shim & Yang, 2004) and Lucene Korean Analyzer (Białecki, Muir, & Ingersoll, 2012), from the collected documents, were used for more robust performance. As a result, 141,509 unique words were extracted. Words appearing fewer than five times and used by less than three authors were filtered out for noise removal. Average length of the mobile text data across all users was 89,188 Korean characters, and the lengths for female and male users were 53,432 characters and 124,945 characters, respectively.

It is noticeable that the number of words exclusively appearing in male documents was 23,568 which was much greater than that of female documents, 3,776. In contrast, the average lengths of both web documents and mobile text data of females were much longer than those of males. Females used about 20% more words than males in case of web documents and two times more characters in case of mobile text data.

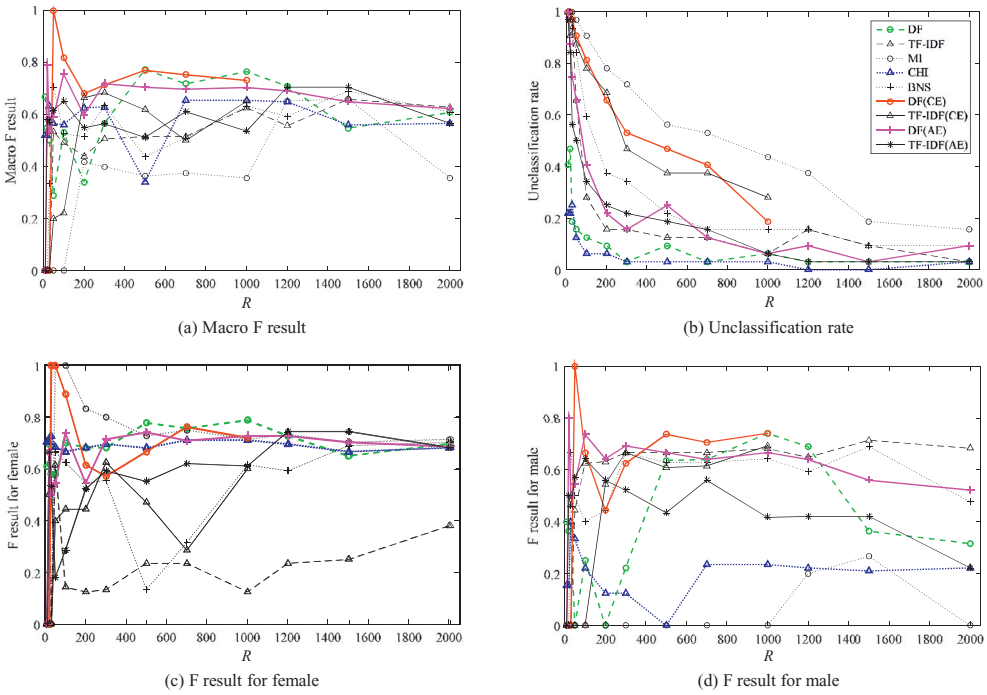


Fig. 2. Prediction performances of the representative wordsets constructed by the nine word evaluation measures according to R in terms of macro F, unclassification rate, and F results for female and male.

4.2. Experiment settings

We conducted four experiments for performance comparisons. The purposes of the experiments are as follows: (1) to show the effectiveness of the proposed framework compared to the conventional ones in terms of time complexity, accuracy, and unclassification rate, (2) to investigate and compare the performances of the nine measures and the four weighting schemes in order to find out which characteristics of a word, among popularity, specificity, and discriminability, are effective for on-device gender prediction, and (3) to observe the performances of the best measure and weighting scheme for the on-device setting. Firstly, Table 6 is presented to show the effectiveness of the proposed framework compared to a conventional method. Then, Figs. 2 and 3 are provided to show the comparative results of the nine measures and the four weighting schemes, respectively. Lastly, Table 7 is newly added to observe the prediction performances of the best method.

Note that the maximum possible number of words in representative wordsets constructed by using CE was set to 1000 since about 1000 distinct words were appearing in the documents by female authors after the noise removal. Threshold ρ in Eqs. (11) and (12) was set to 0.0005 throughout the experiments to make ρ close to the ratio of the average DF across the highly ranked words for a gender by MI to the total number of documents by the gender so that AE considers discriminability close to MI.

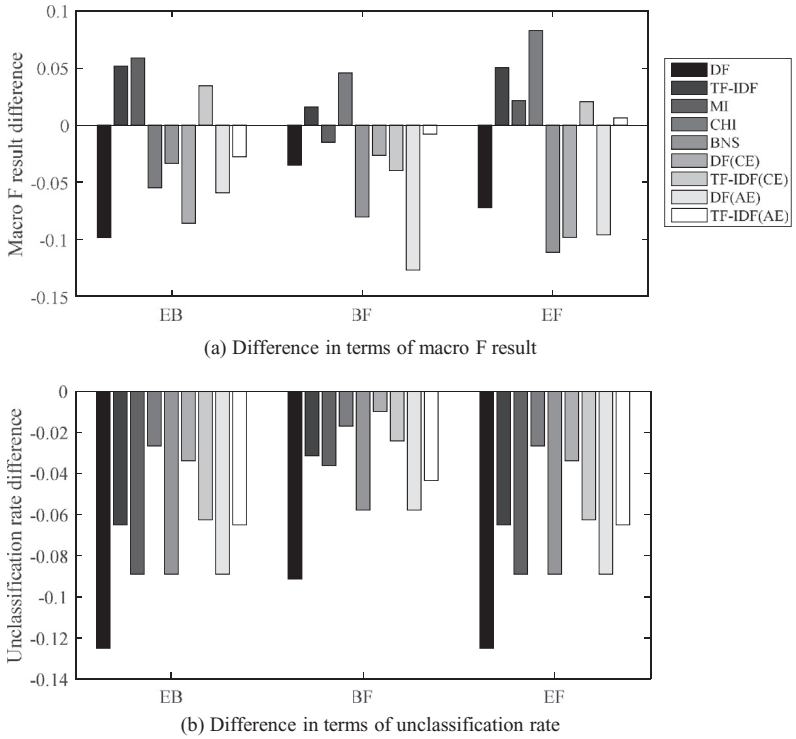
As a conventional gender prediction method, SVM (support vector machine) (Cortes & Vapnik, 1995) which is known for its promising performances in gender prediction problem (Huang et al., 2014) was implemented. Specifically, an SVM was trained by use of the collected web documents with gender labels. Then, genders of the collected mobile text data were predicted by the trained SVM. The documents and the mobile text data were represented as word vectors using unique words from the union of representative wordsets for females and males. Linear kernel was adopted, and the parameters for SVM were set to default values provided in LIBSVM (Chang & Lin, 2011).

As evaluation metrics, F score, macro precision and recall, unclassification rate, and prediction time were considered. Two types of F score were employed: F results of female and male to observe the prediction performances of each gender and macro F to observe the overall performances. F score of a gender is defined as the harmonic mean of precision and recall for the gender, and macro F is the harmonic mean of macro precision and macro recall, which respectively are the weighted arithmetic means of precision and recall for a gender. In addition, there may exist an instance that is classified as neither male nor female because no word in the representative wordsets matches the mobile text data or the matching results of two genders are even. In this regard, unclassification rate was employed, which is the ratio of the number of unclassified instances to the total number of instances. Finally, computation time for prediction was used to evaluate the computational complexity of methods.

4.3. Representative wordsets for genders

In this section, the words selected for the representative wordsets of two genders are presented. Table 4 shows the top five words which were most representative for female according to the measures, while Table 5 shows the top five words for male. Each word is presented in Korean characters along with a description, evaluation score, word length in Korean character, overall TF and DF, and female and male DF.

From Tables 4 and 5, it appears that the representative words for female were related to topics such as cosmetics, plants, fashion, baby, and interior, while those for male were related to game, movie, military, and animation. Furthermore, CHI, MI, and BNS scores for female words were higher than those for male words. These measures consider the discriminability of a word, and the



(a) Difference in terms of macro F result

(b) Difference in terms of unclassification rate

Fig. 3. Performance difference of EB, BF, and EF from BB when BB is a baseline in terms of macro F and unclassification rate.

difference suggests that the words in the female wordsets have higher discriminative power than those in the male wordsets. In case of DF, similar words were selected for both female and male, but while 'we' and 'mind' were ranked high in the female wordset, 'because' appeared in high rank only in the male wordset.

4.4. Performance comparison results

We compared the SVM with the proposed on-device framework using DF(CE) in order to confirm the computational efficiency and prediction effectiveness of the proposed framework. Two types of word matching schemes, binary matching and frequency counting, were compared for the complexity comparison. Table 6 presents the prediction performances of the SVM and the considered on-device methods, averaged across R's, except for cases where R is under 50 as wordsets of these cases were so small that abnormal performances were obtained.

Overall, on-device methods showed similar or slightly better accuracies while required much lower computations. DF(CE) with binary matching outperformed the rest in terms of macro F, but it also showed the highest unclassification rate. Time taken for the prediction by the proposed on-device methods was much smaller than that of the conventional method which requires to build a bag of word for prediction. Moreover, frequency counting consumed about ten to twenty times heavier computation than binary matching, but still it was much lighter than that of the conventional method.

The gender prediction performances for the wordsets constructed by the nine measures were compared by varying R in terms of macro F, unclassification rate, and F results, as shown in Fig. 2. Prediction performances appeared to converge, while unclassification rates decreased as R increased.

As shown in Figs. 2(a) and (b), CHI, which worked best in the previous research on gender prediction using web documents (Miller et al., 2012), showed less competitive performances, implying that the characteristics required for representative words in on-device gender prediction differ from those used in conventional prediction. On the other hand, DF(CE) outperformed the others with an average of macro F of 0.8 when R is between 50 and 1000. However, its unclassification rate ranged between 0.4 and 0.6, which was relatively high compared to those of the others. In addition, when R is between 500 and 1200, DF showed superior performance to the rest with the lowest unclassification rate.

While utilizing words with higher discriminative power resulted in more accurate prediction, it also resulted in higher unclassification rate. A word with high discriminative power but low popularity is more likely unavailable in mobile text data than in web documents as the number of words in mobile text data is much smaller than that in web documents, resulting in higher unclassification rate. This suggests that the popularity of a word should be considered for on-device gender prediction as well as the discriminative power. However, simply combining those two characteristics did not promise enhanced prediction. Although DF(AE) considers more popular words to be more representative than DF(CE) and more discriminative word to be more representative than DF, it did not achieve any improvement on performance compared to both DF(CE) and DF.

From Figs. 2(c) and (d), in which prediction results were evaluated in terms of F values of female and male, it can be concluded that prediction performances for females, with an average F value of 0.7, were better than those of males, with an average F value of 0.5. Similar phenomenon was observed for the evaluation scores of the representative words in Tables 4 and 5, where evaluation scores of female words were higher than those of male words for measures considering discriminability of a word. Particularly, DF performed well for females but poor for males, while TF-IDF which measures topic specificity of a word showed good performances for males but showed the worst performances for females. This implies that topic specific words were effective for predicting male, but as their appearances were rare in mobile text data, prediction performance for males degraded. These results correspond to the conclusion made by Igarashi et al. (2005) that females tend

Table 4
Five most representative words in Korean for females according to the nine word evaluation measures as well as their meaning, evaluation score, word length, TF, DF, and female and male DFs.

Measure	Word	Meaning	Score	Length	TF	DF	female DF	male DF
DF	생각	Thinking	0.3796	2	164105	63847	20264	43583
	사람	Person	0.3271	2	165161	56754	17459	39295
	우리	We	0.3221	2	117453	44386	17192	27194
	사진	Photo	0.2716	2	114246	39831	14501	25330
	마음	Mind	0.249	2	75798	37041	13290	23751
TF-IDF	아티브	A brand name of laptop produced by Samsung	0.2125	3	41	4	1	3
	평화누리길	A tracking course in Korea	0.2067	5	41	4	1	3
	채피	Chappie (a movie title)	0.1895	2	136	19	1	18
	조가비	Shell	0.1793	3	374	45	9	36
	한련화	Garden nasturtium	0.1723	3	169	7	5	2
MI	베네피트	Benefit (a cosmetic brand)	1.2649	4	62	32	32	0
	데일리백	Daily bag	1.2649	4	51	27	27	0
	짜장떡볶이	Rice cake with black soybean sauce	1.2649	5	6	4	4	0
	아기옷	Baby clothes	1.2649	3	29	20	20	0
	정리대	Arrangement shelf	1.2649	3	8	6	6	0
CHI	엄마	Mom	8128.6041	2	32019	11113	7287	3826
	언니	Women's older sister	6119.0877	2	8642	3601	3109	492
	완전	Completely	4782.6742	2	17214	10775	6179	4596
	신랑	Husband	4638.5512	2	4730	2291	2105	186
	아이	Child	4158.4591	2	59673	18937	9134	9803
BNS	배송정보	Delivery information	1.7259	4	1021	245	244	1
	복용법	Intake method	1.4752	3	128	115	114	1
	도시농부	Urban farmer	1.4676	4	560	292	288	4
	새언니	Women's sister-in-law	1.4581	3	156	109	108	1
	블렌더	Blender	1.4522	3	131	107	106	1
CE(DF)	세부퍼시픽	Cebu Pacific (a Philippines airline)	0.0019	5	360	99	99	0
	로즈제라늄	Rose geranium	0.0016	5	318	88	88	0
	난리난리	Overly expressing a fuss	0.0014	4	82	77	77	0
	딸램	A cute word for calling a daughter	0.0014	2	148	75	75	0
	육아맘	Child-rearing mother	0.0011	3	63	59	59	0
CE(TF-IDF)	곱창고	A restaurant in Korea	0.0416	3	29	4	4	0
	로제스파게티	Rose sauce spaghetti	0.0402	6	35	5	5	0
	싸롱	Salon	0.0402	2	28	4	4	0
	글리터	Glitter	0.0345	3	36	6	6	0
	온수매트	Hot water mat	0.0332	4	52	9	9	0
AE(DF)	리폼	Reform	0.0094	2	1996	565	501	64
	택배비	Delivery fee	0.0090	3	718	527	481	46
	시댁	Women's family-in-law	0.0089	2	715	520	476	44
	수납장	Storage cabinet	0.0086	3	1007	509	460	49
	패브릭	Fabric	0.0085	3	1042	519	455	64
AE(TF-IDF)	아티브	A brand name of laptop produced by Samsung	0.2125	3	41	4	1	3
	평화누리길	A tracking course in Korea	0.2067	5	41	4	1	3
	한련화	Garden nasturtium	0.1723	3	169	7	5	2
	엘레노어	Eleanor	0.1091	4	39	7	1	6
	크리스틴	Christine	0.0976	4	45	4	2	2

to talk about general topics with other people, while males talk about specific topics such as sports and games.

Figs. 3(a) and (b) show the effects of employing different types of similarities mentioned in Section 3.4.2. Specifically, the bars indicate the performance differences between EB, BF, and DF compared to BB. In terms of macro F, the overall performances of TF-IDF, MI, CHI, TF-IDF(CE), and TF-IDF(AE) increased when $\vec{s_g}$ or $\vec{f_{gu}}$ was utilized while those of DF, BNS, DF(CE), and DF(AE) decreased. Using $\vec{s_g}$ for DF, BNS, DF(CE), and DF(AE) was not effective because the word evaluation scores were highly correlated to the popularity of a word rather than its discriminative power for genders.

Moreover, unclassification rates of all measures decreased when $\vec{s_g}$ or $\vec{f_{gu}}$ was employed as shown in Fig. 3(b). Particularly, the differences were significant for DF since there exist a large number of common words between female and male wordsets, resulting in many even matches when BB was utilized. From the results of EB and BF, utilizing $\vec{s_g}$ turned out to be more effective than employing $\vec{f_{gu}}$ in terms of unclassification rate.

Lastly, we have examined the gender prediction performances of the two best measures, DF(CE) and DF, according to the various R, as shown in Table 7. Particularly, prediction performances of random prediction were additionally provided to observe the expected performances when the unclassified cases are randomly determined. Note that 'random' indicates random prediction with 50% probability for unclassified cases while 'None' indicates no post processing for the cases.

Overall, maximum macro F was achieved by DF(CE) with the value of 0.87 for 'None', while it was acheived by DF with the value of 0.75 for 'random'. As R increased, unclassification rate decreased for 'None', and Macro F increased for 'Random'. In addition, macro precision values were larger compared to macro recall values, implying that the proposed framework was able to predict the gender with clear evidence, but there may exist some subjects who were outliers.

5. Conclusion

In this paper, we proposed an on-device gender prediction framework using representative wordsets. A representative wordset of a gender is constructed by evaluating gender representativeness

Table 5
Five most representative words in Korean for males according to the nine word evaluation measures as well as their meaning, evaluation score, word length, TF, DF, and female and male DFs.

Measure	Word	Meaning	Score	Length	TF	DF	female DF	male DF
DF	생각	Thinking	0.3211	2	164105	63847	20264	43583
	사람	Person	0.2895	2	165161	56754	17459	39295
	정도	Degree	0.2338	2	91820	45761	14026	31735
	때문	Because	0.2236	2	86371	42470	12115	30355
	우리	We	0.2003	2	117453	44386	17192	27194
TF-IDF	홍대데이트	Dating in Hongdae (a trendy place in Korea)	0.3925	5	74	6	5	1
	금녀	Prohibiting female	0.1487	2	80	5	2	3
	빈스빈스	A cafe in Korea	0.1249	4	46	7	6	1
	귀찬	Being annoyed	0.1145	2	83	8	4	4
	직쏘	Jigsaw (a movie character)	0.1142	2	194	11	1	10
MI	내수면	A township in Korea	0.3316	3	16	9	0	9
	시험비행	Test flight	0.3316	4	188	125	0	125
	축구스타	Soccer star player	0.3316	4	10	8	0	8
	청룡기	A baseball championship in Korea	0.3316	3	10	8	0	8
	배관공	Plumber	0.3316	3	12	10	0	10
CHI	감독	Director	2961.146	2	37607	13196	1011	12185
	영화	Movie	2710.3639	2	155832	24593	3514	21079
	액션	Action	1244.4212	2	12768	4801	269	4532
	게임	Game	1219.245	2	40965	8371	957	7414
	전투	Battle	1155.5253	2	13629	3984	169	3815
BNS	사활	Life or death	1.5936	2	5748	2939	8	2931
	보병	Infantry	1.5438	2	2173	674	1	673
	코믹북	Comic book	1.4992	3	863	592	1	591
	록음악	Rock music	1.4387	3	1129	495	1	494
	항공모함	Aircraft carrier	1.4110	4	1994	456	1	455
CE(DF)	구축함	Destroyer (a battleship)	0.0046	3	3302	628	0	628
	미해군	US Navy	0.0034	3	1155	466	0	466
	불펜	Bull pen	0.0029	2	1117	392	0	392
	스타팅	Starting	0.0027	3	1593	370	0	370
	코믹콘	Comic Con	0.0023	3	454	316	0	316
CE(TF-IDF)	이륜자동차	Motorcycle	0.0743	5	425	34	0	34
	미드웨이	Midway Islands	0.0673	4	917	81	0	81
	수아레즈	Suarez (a soccer player)	0.0615	4	62	6	0	6
	부동산가격	Real estate price	0.0575	5	87	9	0	9
	북알프스	North of the Alps	0.0476	4	144	18	0	18
AE(DF)	사활	Life or death	0.0216	2	5748	2939	8	2931
	하늘소	Long horned beetle	0.0144	3	2688	1971	10	1961
	포켓몬스터	Pocket Monster (a cartoon)	0.0119	5	6321	1623	7	1616
	전투기	Combat plane	0.0103	3	6711	1421	19	1402
	파워포인트	MS PowerPoint	0.0091	5	5407	1249	10	1239
AE(TF-IDF)	금녀	Prohibiting female	0.1487	2	80	5	2	3
	직쏘	Jigsaw (a movie character)	0.1142	2	194	11	1	10
	타잔	Tarzan (a movie character)	0.0983	2	290	18	1	17
	논개	A historical figure in Korea	0.0932	2	68	5	2	3
	연아	Yuna Kim (a Korean figure skater)	0.0852	2	45	4	1	3

Table 6
Performances of the conventional method, SVM, and on-device methods, DF(CE), in terms of macro F, unclassification rate, and prediction time in seconds.

Methods	Conventional	On-device	
	SVM (linear)	DF(CE)	
		Binary matching	Frequency counting
Macro F value	0.75	0.78	0.73
Unclassification rate	0.33	0.57	0.55
Prediction time (s)	845.72	0.67	8.73

of words extracted from web documents. Inside a mobile device, the text data of a user is matched to the representative words to measure the similarities between the text data and the wordset of each gender. Then, a user's gender is determined as the one with higher similarity. Nine word evaluation measures, DF, TF-IDF, MI, CHI, BNS, CE(DF), CE(TF-IDF), AE(DF), and AE(TF-IDF) were considered, and their gender prediction performances were compared.

From the experiment results, five conclusions are drawn. First, males used more topic-specific words than females as concluded

Table 7
Gender prediction performances of the two best measures, DF(CE) and DF, with or without the random prediction for the unclassified cases in terms of macro F, macro precision, macro recall, and unclassification rate.

Measure	Unclassified case	Metric	*R*		
			100	500	1000
DF(CE)	None	Macro F	0.87	0.76	0.73
		Macro precision	0.92	0.83	0.73
		Macro recall	0.83	0.71	0.73
		Unclassified rate	0.81	0.47	0.19
	Random	Macro F	0.56	0.60	0.69
		Macro precision	0.56	0.60	0.69
		Macro recall	0.56	0.59	0.69
		Unclassified rate	0.00	0.00	0.00
DF	None	Macro F	0.56	0.77	0.77
		Macro precision	0.55	0.82	0.77
		Macro recall	0.57	0.72	0.77
		Unclassified rate	0.13	0.09	0.06
	Random	Macro F	0.58	0.74	0.75
		Macro precision	0.60	0.77	0.75
		Macro recall	0.56	0.72	0.75
		Unclassified rate	0.00	0.00	0.00

by Igarashi et al. (2005). This was observed from the difference between the numbers of words only used by females and males in Table 3. Second, the average length of documents written by females was longer than that by males as in Table 3. Third, in addition to the discriminative power, the popularity of a word was important for on-device gender prediction since mobile text data is relatively short and composed of a limited number of words compared to web documents. Fourth, gender prediction performances of females were better than those of males. It can be explained from the first and second findings that the probability of word matching between mobile text data and representative wordset of females is higher than that of males as females write longer documents with more general words. Finally, with respect to the word score type for the similarity calculation, employing word evaluation scores of TF-IDF, MI, CHI, TF-IDF(CE), and TF-IDF(AE), which are measuring the discriminability of words, has improved gender prediction performances compared to employing binary scores, while those of DF, BNS, DF(CE), and DF(AE), which are measuring the popularity, have degraded the performances.

The contributions of the paper are summarized into three points. First, by adopting the on-device analytics concept from expert and intelligent systems, the proposed framework was able to accurately predict the gender of a mobile device user by using textual data without a privacy concern. Moreover, the simple word matching scheme enabled the framework to operate with limited resources of a mobile device. Second, from the experiments conducted on real-world dataset acquired from actual blogs and mobile devices, the potential of the proposed framework was confirmed, important characteristics for the representative words were investigated, and the trade-offs between diverse parameters and performance evaluation measures were observed. Third, the proposed framework can be extended to address similar privacy issue in expert and intelligent systems with a little modification.

The strengths and weaknesses of the proposed framework are as follows. The proposed framework is simple yet very powerful approach. It utilizes the high gender discriminability of user generated text without risking privacy of users, and requires low computational power that it is possible to embed the framework into mobile devices. In addition, the proposed framework is language independent since it does not utilize language-dependent features. Therefore, it can be easily extended to other languages with a small modification such as stop word during text processing. However, the proposed framework has a weakness that there exist unclassified cases according to the quality of representative wordsets. Moreover, for some language, where the differences between texts used by females and males are subtle, the prediction performance can be degraded.

For the future work, we plan to collect more diverse web documents such as SNSs and web community services, and it will facilitate better gender prediction performances since those services are closely related to users' feeling and their everyday lives that their contents are more similar to the contents of mobile text data than blog documents. Moreover, it will be very interesting to investigate the relevance between language properties and the gender prediction performance using the proposed framework. The authors are currently working on the development of multi-phase representative word matching to resolve the problem of unclassification, which can preserve the accurate prediction using discriminative words while minimize the number of unclassified cases by performing matching with more popular words, recursively. Lastly, with the advances in algorithms and computing power of mobile devices, the conventional methods for semantic word search can be implemented in small mobile devices in recent future.

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. 2013R1A2A2A03013947).

References

Aizawa, A. (2003). An information-theoretic perspective of TF-IDF measures. *Information Processing and Management, 39*(1), 45–65.
Alowibdi, J. S., Buy, U. A., & Yu, P. (2013a). Empirical evaluation of profile characteristics for gender classification on twitter. In *Proceedings of the IEEE international conference on machine learning and applications: 1* (pp. 365–369).
Alowibdi, J. S., Buy, U. A., & Yu, P. (2013b). Language independent gender classification on twitter. In *Proceedings of the IEEE/ACM international conference on advances in Social Networks Analysis and Mining* (pp. 739–743).
Azcarraga, A., Liu, M. D., & Setiono, R. (2012). Keyword extraction using backpropagation neural networks and rule extraction. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 1–7).
Barragáns-Martínez, B., Costa-Montenegro, E., & Juncal-Martínez, J. (2015). Developing a recommender system in a consumer electronic device. *Expert Systems with Applications, 42*(9), 4216–4228.
Białecki, A., Muir, R., & Ingersoll, G. (2012). Apache lucene 4. In *Proceedings of the workshop on open source information retrieval* (pp. 17–24).
Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27.
Cheng, N., Chandramouli, R., & Subbalakshmi, K. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78–88.
Choi, I. Y., Oh, M. G., Kim, J. K., & Ryu, Y. U. (2016). Collaborative filtering with facial expressions for online video recommendation. *International Journal of Information Management, 36*(3), 397–402.
Chong, A. Y.-L. (2013). Predicting m-commerce adoption determinants: a neural network approach. *Expert Systems with Applications, 40*(2), 523–530.
Chung, W. (2014). Bizpro: extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management, 34*(2), 272–284.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.
Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W. (2012). Author gender prediction in an email stream using neural networks. *Journal of Intelligent Learning Systems and Applications, 4*(3), 169–175.
Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the ACM international conference on Knowledge discovery and data mining* (pp. 15–24).
Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research, 3*, 1289–1305.
Hajmohammadi, M. S., Ibrahim, R., & Selamat, A. (2014). Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. *Engineering Applications of Artificial Intelligence, 36*, 195–203.
Han, S., Shen, H., Philipose, M., Agarwal, S., Wolman, A., & Krishnamurthy, A. (2016). MCDNN: an approximation-based execution framework for deep stream processing under resource constraints. In *Proceedings of the international conference on mobile systems, applications, and services* (pp. 123–136).
Huang, F., Li, C., & Lin, L. (2014). Identifying gender of microblog users based on message mining. *Lecture Notes in Computer Science, 8485*, 488–493.
Igarashi, T., Takai, J., & Yoshida, T. (2005). Gender differences in social network development via mobile phone text messages: a longitudinal study. *Journal of Social and Personal Relationships, 22*(5), 691–713.
Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems, 51*, 35–47.
Jeong, S., Cheng, D., Song, H., & Kalasapur, S. (2009). Non-collaborative interest mining for personal devices. In *Proceedings of the IEEE symposium on computational intelligence and data mining* (pp. 179–186).
Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2006). Chat mining for gender prediction. *Lecture Notes in Computer Science, 4243*, 274–283.
Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on multi-source multilingual information extraction and summarization* (pp. 17–24).
López-Monroy, A. P., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015). Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems, 89*, 134–147.
Loscalzo, S., Wright, R., & Yu, L. (2014). Predictive feature selection for genetic policy search. *Autonomous Agents and Multi-Agent Systems*, 1–33.
Lu, S.-H., Chiang, D.-A., Keh, H.-C., & Huang, H.-H. (2010). Chinese text classification by the Naïve bayes classifier and the associative classifier with multiple confidence threshold values. *Knowledge-based systems, 23*(6), 598–604.
Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on twitter using stream algorithms with N-gram character features. *International Journal of Intelligence Science, 2*(4), 143–148.

Mukherji, A., Srinivasan, V., & Welbourne, E. (2014). Adding intelligence to your mobile device via on-device sequential pattern mining. In *Proceedings of the ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 1005–1014).

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications, 42*(9), 4348–4360.

Nunes, I., Miles, S., Luck, M., Barbosa, S., & Lucena, C. (2015). Decision making with natural language based preferences and psychology-inspired heuristics. *Engineering Applications of Artificial Intelligence, 42*, 16–35.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the ACM international workshop on search and mining user-generated contents* (pp. 37–44).

Rogati, M., & Yang, Y. (2002). High-performing feature selection for text classification. In *Proceedings of the international conference on information and knowledge management* (pp. 659–661).

Roul, R. K., Devanand, O. R., & Sahay, S. K. (2014). Web document clustering and ranking using tf-idf based apriori approach. In *Proceedings of the ieee international conference on advances in computer engineering and applications, 2* (p. 34).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Seneviratne, S., Seneviratne, A., Mohapatra, P., & Mahanti, A. (2014). Your installed apps reveal your gender and more!. *ACM Mobile Computing and Communications Review, 18*(3), 55–61.

Shim, K., & Yang, J. (2004). High speed Korean morphological analysis based on adjacency condition check. *Journal of Korea Information Science Society: Software and Applications, 31*(1), 88–99.

Shotton, J., Sharp, T., Kohli, P., Nowozin, S., Winn, J., & Criminisi, A. (2013). Decision jungles: compact and rich models for classification. In *Proceeding of the advances in neural information processing systems* (pp. 234–242).

Srinivasan, V., Moghaddam, S., Mukherji, A., Rachuri, K., Xu, C., & Tapia, E. M. (2014). On-device mining of mobile users' co-occurrence patterns. In *Proceedings of the international workshop on mobile computing systems and applications*.

Wang, Y., Liu, Y., Feng, L., & Zhu, X. (2015). Novel feature selection method based on harmony search for email classification. *Knowledge-Based Systems, 73*, 311–323.

Weiss, G. M., & Lockhart, J. W. (2011). Identifying user traits by mining smart phone accelerometer data. In *Proceedings of the ACM international workshop on knowledge discovery from sensor data* (pp. 61–69).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the international conference on machine learning: 97* (pp. 412–420).

Yeh, C.-H., Wang, Y.-S., & Yieh, K. (2016). Predicting smartphone brand loyalty: consumer value and consumer-brand identification perspectives. *International Journal of Information Management, 36*(3), 245–257.

Ying, J. J.-C., Chang, Y.-J., Huang, C.-M., & Tseng, V. S. (2012). Demographic prediction based on users mobile behaviors. In *Proceedings of the mobile data challenge workshop*.

Zhong, E., Tan, B., Mo, K., & Yang, Q. (2013). User demographics prediction based on mobile data. *Pervasive and Mobile Computing, 9*(6), 823–837.