Article

Machine-Learning-Based Gender Distribution Prediction from Anonymous News Comments: The Case of Korean News Portal

Jong Hwan Suh

Department of Management Information Systems & BERI, Gyeongsang National University, 501 Jinjudae-ro, Jinju-si 52828, Korea; jonghwan.suh@gnu.ac.kr; Tel.: +82-55-772-1537; Fax: +82-55-772-1539

Abstract: Anonymous news comment data from a news portal in South Korea, naver.com, can help conduct gender research and resolve related issues for sustainable societies. Nevertheless, only a small portion of gender information (i.e., gender distribution) is open to the public, and therefore, it has rarely been considered for gender research. Hence, this paper aims to resolve the matter of incomplete gender information and make the anonymous news comment data usable for gender research as new social media big data. This paper proposes a machine-learning-based approach for predicting the gender distribution (i.e., male and female rates) of anonymous news commenters for a news article. Initially, the big data of news articles and their anonymous news comments were collected and divided into labeled and unlabeled datasets (i.e., with and without gender information). The word2vec approach was employed to represent a news article by the characteristics of the news comments. Then, using the labeled dataset, various prediction techniques were evaluated for predicting the gender distribution of anonymous news commenters for a labeled news article. As a result, the neural network was selected as the best prediction technique, and it could accurately predict the gender distribution of anonymous news commenters of the labeled news article. Thus, this study showed that a machine-learning-based approach can overcome the incomplete gender information problem of anonymous social media users. Moreover, when the gender distributions of the unlabeled news articles were predicted using the best neural network model, trained with the labeled dataset, their distribution turned out different from the labeled news articles. The result indicates that using only the labeled dataset for gender research can result in misleading findings and distorted conclusions. The predicted gender distributions for the unlabeled news articles can help to better understand anonymous news commenters as humans for sustainable societies. Eventually, this study provides a new way for data-driven computational social science with incomplete and anonymous social media big data.

Keywords: anonymity; social media; big data; news comments; gender prediction; word embedding; machine learning

check for updates

Citation: Suh, J.H. Machine-Learning-Based Gender Distribution Prediction from Anonymous News Comments: The Case of Korean News Portal. *Sustainability* **2022**, *14*, 9939. https://doi.org/10.3390/su14169939

Academic Editors: Miltiadis D. Lytras and Andreea Claudia Serban

Received: 18 July 2022
Accepted: 8 August 2022
Published: 11 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Purpose

Social media provide a venue for individuals to engage in online activities and interact with others [1,2]. On this wise, the advent and extension of social media have been producing big data, and they are undoubtedly leading the era of big data [2,3]. Moreover, with the COVID-19 outbreak, social media has become the platform of more choices for public opinions, perceptions, and attitudes toward various events, including public health policies [4].

Now the developed capacity to collect and analyze such social media big data provides unprecedented opportunities for social science research areas, whose primary interest has been human dynamics [3,5]. Specifically, gender has been continuously receiving attention in prior research works to understand human dynamics through social media for sustainable societies [6,7]. In addition, since gender information is required to perform all gender research, it is essential to obtain gender information, e.g., gender labels, to study gender and related issues through social media big data for sustainable societies.

However, while it is inevitably difficult to obtain individual gender data [8], the anonymity and privacy policy of social media have made it difficult or impossible to acquire gender information from social media [9]. As a result, most prior gender research with social media has been made by using small-size or large-size data, where gender information could be collected [10,11]. If unavailable, gender information had to be manually annotated by researchers [12], or simple estimation approaches were adopted based on relevant cues such as names [13,14].

Nevertheless, manual labeling cannot be applied to the size of social media big data. In addition, if the social media data are anonymous in accordance with a privacy policy, even a simple estimation approach using names as relevant cues cannot be adopted. Hence, prior gender researchers have not even tried to collect anonymous social media big data whose gender information is incomplete, i.e., partially open to the public, due to anonymity, and it led to a situation of little gender research using various sources of anonymous social media big data.

In the same context, anonymous news comment data of news portals in South Korea, possibly useful for gender research as social media big data, have rarely been considered for gender research and remain unexplored. However, fortunately, one of the Korean news portals, naver.com, provides the gender information of anonymous news commenters for a news article, i.e., male and female rates as gender distribution. Therefore, Lee and Ryu [15] could use naver.com to study gender differences among news categories and sub-topics. However, for each news category, they used only the top 30 most-viewed news articles, whose gender distributions were available, i.e., labeled news articles.

Like this, naver.com has the policy to make the gender distribution for a news article open to the public only if the number of its direct news comments exceeds a specific number, i.e., 100. News comments on a news article in naver.com can be classified into two: (i) news comments that reply directly to the news article and (ii) replies to news comments. In this paper, news comments were used to represent both, while the term 'direct news comment' was adopted to be distinguished from the replies to news comments.

Because of such a restrictive policy, the gender distribution for almost all news articles, which have less than 100 direct news comments, remains unrevealed on naver.com, and therefore, those news articles can be considered unlabeled, i.e., unlabeled news articles. This study also found that in around 90% of news articles published on naver.com for two months, from 6/1 to 7/31 in 2018, the gender distribution of news commenters of a news article is not disclosed. Therefore, to use all the anonymous news comment data in naver.com as social media big data for gender research, it is necessary and valuable to think of a method to predict the gender distribution of anonymous new commenters for a news article.

In these circumstances, focusing on the Korean news portal, naver.com, this paper aimed to enable its anonymous news comments to be used as social media big data for gender research. To do so, this paper proposed a machine-learning-based method for predicting the gender distribution of anonymous news commenters for a news article, represented by the characteristics based on the news comments of the news article. Using the labeled news articles, the proposed method was evaluated with different prediction techniques, and the best prediction technique was explored and selected. In addition, for the unlabeled news articles, their unknown gender distributions were predicted using the chosen best prediction technique and the labeled news articles. The predicted gender distributions for the unlabeled news articles were explored and compared with the known gender distributions for the labeled news articles.

1.2. Reviews on Related Works

Table 1 describes the recent works on gender research that used social media, and they can be summarized from various aspects such as purpose, data, and the used gender information. The findings in Table 1 can be summarized as follows:

First, the purposes of recent works on gender research that used social media can be divided into two: (i) gender prediction and (ii) gender analysis with the collected or predicted gender information, but most of the prior works have aimed at gender prediction.

Second, in terms of data, various data sources have been used for gender prediction, but news articles and their comments have rarely been considered so far. In addition, the social media data of prior research have come from many kinds of languages, but Korean social media data have rarely been studied for gender prediction. The data size used by the previous works in Table 1 varied from ten units to million units but did not exceed millions of units.

Third, regarding the used gender information, most previous works have targeted user gender, binary as male or female. Still, no prior work has been conducted on gender distribution in user collectives. In cases with gender labels available, small and large data, where gender information could be verified and collected, were used; users without gender information were to be excluded from the initially collected data [10,13,16].

Fourth, on the other hand, if gender labels were not available, gender labels were manually annotated by researchers, or simple estimation approaches were adopted based on related cues, e.g., name, even though they could be unreliable because they discarded users whose gender could not be ascertained (e.g., neutral names). However, such manual annotation is inappropriate for big data, and simple estimation approaches have an increased risk of inaccurate estimates. Though the name-based estimation for gender prediction has been adopted frequently, it cannot be applied to anonymous social media data such as in this study.

Thus, to overcome the problem of incomplete gender information, gender prediction using machine learning has gained much attention from recent gender research on social media. Regarding the machine-learning-based labeling for gender, prior studies shown in Table 1 can be classified by their data representations and prediction techniques, and it is as shown in Table 2.

In detail, according to their data sources, prior gender research from social media, shown in Table 1, has used various features for machine-learning-based gender prediction from social media. Recently, unstructured data such as images, voices, and textual data have been the features of their interest.

Particularly to represent gender by using textual data, most prior works have adopted the vector space model based on linguistic features such as a bag of words (BOW). Previous approaches to extracting linguistic features can be divided into closed and open vocabularies. While the prior works using the open vocabulary disregarded less informative linguistic features [23], gender could be differentiated by the open vocabulary in social media [24]. Moreover, recently, a few works have started adopting word embeddings. Specifically, the word2vec approach for word embeddings could include and use all linguistic features without disregarding linguistic features, and therefore, word embeddings were used to generate sentence embeddings [25].

Most of the prior studies in Table 1, which used machine-learning labeling approaches to identify gender from social media, have used various prediction techniques. They are summarized as shown in Table 2. Because most of the prior works considered gender as binary, the prediction techniques used for gender prediction in social media were mostly classifiers such as logistic regression (LR), decision tree (DT), and support vector machine (SVM).

Table 1. The recent works on gender research from social media.

Previous Work	Purpose	Data			Gender Information Used			
		Source	**Language**	**Data Size**	**Target**	**Type**	**Label Availability**	**Labeling Method**
Cheng, Chandramouli and Subbalakshmi [13]	Gender prediction	Reuters newsgroup dataset and Enron e-mail dataset	English	6769 news messages and 8970 e-mails	User	Binary	No	Estimated from name
Otterbacher [16]	Gender prediction	The internet movie database (IMDb)	English	31,300 reviews	User	Binary	Yes	Used the male/female filter to collect reviews with gender information
Bamman, et al. [17]	Gender analysis (gender was estimated and used for studying the impact of gender on both linguistic style and social networks in social media text)	Twitter	English	9,212,118 tweets of 14,000 users	User	Binary	No	Estimated from name
Choi, et al. [18]	Gender prediction	Blog posts	Korean	189,127 web documents and mobile text data from 32 users	User	Binary	Yes	Collected
Hosseini and Tammimy [19]	Gender prediction	News comments of LA Times	English	30 comments	User	Binary	Yes	Collected
Teso, et al. [20]	Gender analysis (gender prediction was used to check if the language of online communication is gender-free or influenced by patterns of male dominance)	Online consumer reviews (www.ciao.co.uk)	English	2083 reviews from each gender	User	Binary	Yes	Collected
Al-Ghadir and Azmi [21]	Gender prediction	Social forum in Saudi Arabia (www.eqla3.com)	Arabic	433,199 posts of 100,000 randomly selected topics	User	Binary	Yes	Collected
Hirt, Kühl and Satzger [6]	Gender prediction	Tweets, Twitter user name, and Twitter profile image	German	2916 profiles	User	Binary	Yes	Collected
Hussein, et al. [22]	Gender prediction	Tweets	Egyptian	70 accounts for each gender and 1000 tweets for each account for the balanced dataset	User	Binary	Yes	Collected

Table 1. *Cont.*

Previous Work	Purpose	Data			Gender Information Used			
		Source	Language	Data Size	Target	Type	Label Availability	Labeling Method
Kucukyilmaz, Deniz and Kiziloz [10]	Gender prediction	Chat messages in Heaven BBS	Turkish	835 users	User	Binary	Yes	Collected
López-Monroy, González and Solorio [11]	Gender prediction	Tweets	English and Spanish	3600 English users and 4300 Spanish users from PAN-CLEF 2017	User	Binary	Yes	Collected
López-Santillán, Montes-Y-Gómez, González-Gurrola, Ramírez-Alonso and Prieto-Ordaz [7]	Gender prediction (as well as predictions on other user characteristics such as age and personality traits)	Social media outlets, blogs, Twitter, and hotel reviews	English, Arabic, Spanish, Portuguese, Dutch, Italian	PAN-CLEF since 2013	User	Binary	Yes	Collected
Aman, Smith-Colin and Zhang [14]	Gender analysis (gender was estimated to investigate how satisfaction levels expressed in the reviews vary across gender)	Rider-generated reviews on the Google Play Store and Apple App Store	English	12,026 rider-generated reviews	User	Binary	No	Estimated from name
Das and Paik [12]	Gender prediction	Names from news reports (CoNLL-g and IE-ER-g), Wikipedia articles (Wiki-g), and textbooks (Textbook-g)	English	8650 names	User	Binary	Yes	Manually annotated
This study	Gender prediction	News portal in South Korea (naver.com)	Korean	177,735 news articles and 14,896,043 anonymous news comments	Collective users (news commenters for a news article)	Continuous (male and female rates as gender distribution)	Partially available	Collected if available

Table 2. Prior works on machine-learning-based gender prediction from social media.

Previous Work	Text Representation	Prediction Technique
Cheng, Chandramouli and Subbalakshmi [13]	Feature-based (character-based features, word-based features, syntactic features, structural features, and function words)	Bayesian logistic regression (BLR), AdaBoost decision tree (ADT), and support vector machine (SVM)
Otterbacher [16]	Feature-based (writing style features, review content features, and review/movie metadata)	Logistic regression (LR)
Bamman, Eisenstein and Schnoebelen [17]	Feature-based (lexical features and network features)	LR
Choi, Kim, Kim, Park and Park [18]	Feature-based (word-based features)	Used textual similarity of a user, estimated from gender-known documents, and SVM as the conventional method
Teso, Olmedilla, Martínez-Torres and Toral [20]	Feature-based (lexical features, sentiment features, and content features)	Naïve Bayes (NB)
Al-Ghadir and Azmi [21]	Feature-based (word-based features)	SVM, *k*-nearest neighbors (*k*-NN)
Hirt, Kühl and Satzger [6]	Feature-based (word-based features, name-based features, and image-based features)	NB, SVM, and random forest (RF) for text classifier, Levenshtein distance for name classifier, the third-party image classifier for image classifier, and finally, meta classifier by integrating results from text, name, and image classifiers as a meta-feature vector
Hussein, Farouk and Hemayed [22]	Feature-based (emoji-based features, female suffix features, and function-based features) and word-embedding-based	RF, NB, and LR
Kucukyilmaz, Deniz and Kiziloz [10]	Feature-based (word-based features)	*k*-NN, NB, SVM, RF, LR, and decision tree (DT)
López-Monroy, González and Solorio [11]	Feature-based (word-based features and meta-word features)	SVM
López-Santillán, Montes-Y-Gómez, González-Gurrola, Ramírez-Alonso and Prieto-Ordaz [7]	Word embedding-based	SVM/support vector regression (SVR), RF/random forest regression (RFR), extra trees (ET)/extra trees regression (ETR), and *k*-NN/*k*-nearest neighbors regression (*k*-NNR)
Aman, Smith-Colin and Zhang [14]	Feature-based (content-based features and sentiment features)	LR
Das and Paik [12]	Feature-based (sequence features)	Transformer network and deep learning methods such as bidirectional long short-term memory network (BiLSTM), the vanilla transformer network, fine-tuned BERT, BERTBiLSTM-CRF, and a unified machine reading comprehension (MRC) framework
This study	Word embedding-based	Multiple linear regression (MLR), decision tree regression (DTR), SVR, *k*-NNR, neural network regression (NNR)

In terms of the taxonomies mentioned above, theoretically, this paper can be classified as shown in Tables 1 and 2. In detail, Table 1 shows that this paper focused on gender prediction as a research purpose. This study eventually extends and contributes to the literature by enabling gender research even with incomplete gender information from anonymous social media big data. Moreover, this paper introduced and used anonymous news comments as a new data source for gender research in social media. Compared to prior works in Table 1, this study introduced news comments written mainly in Korean, and it extends the literature in terms of language diversity.

Unlike prior works in Table 1, when considering the characteristic of gender information of the Korean news portal, naver.com, according to its privacy protection, the gender distribution of news commenters for a news article was chosen as gender information, not binary but continuous between 0 and 1. By doing so, this study provided a new perspective on dealing with gender, i.e., a collective gender. It helps to extend the scope of gender from the individual to the collective level.

Related to data representation in Table 2, this study used news comments as text data to represent the gender distribution of collective users. However, it is hard to find explicit cues for gender prediction in news comments due to the following reasons: (i) most of the news comments are very short, while they vary in length; (ii) they are not grammatically correct in most cases; and (iii) they contain different kinds of unstructured features such as emojis and characters of facial expressions.

Hence, unlike most previous works that used linguistic features for text data representation, this study adopted word embeddings, word2vec, to construct text data representation. In detail, to generate text representation for a news article, this study aggregated the word embeddings of words in the news comments of the news article.

Moreover, in terms of prediction techniques in Table 2, because this study considered the gender distribution of news commenters for a news article as a continuous value between 0 and 1, not binary, prediction techniques commonly used for the continuous target variable in machine-learning applications were used. In addition, their performances were evaluated and compared with each other.

1.3. Research Gaps and Questions

The research gaps, identified from the literature reviews, could be summarized as follows:

First, the previous gender research has rarely paid attention to using news articles and their anonymous news comments as social media big data for both predicting gender information and gender research using the predicted gender information.

Second, it is currently unclear how information technologies can be used to predict the gender distribution of anonymous news commenters for labeled news articles.

Third, it has not been explored how different the predicted gender distributions for unlabeled news articles, which are far more than labeled news articles, will be when compared to the true gender distributions for labeled news articles.

Eventually, considering the findings from the related works and the above-mentioned research gaps, the research questions of this study could be formulated as below:

First, regarding the task of predicting the gender distribution of anonymous news commenters for a labeled news article,

- RQ1. Which prediction technique will be best suited in a statistically significant way? In detail, if 10-fold cross-validation as an experiment is repeated 50 times for each prediction technique, which prediction technique will be best? Will this result be the same as the comparison results by pairwise *t* tests?
- Second, if the best prediction technique is used,
- RQ2. How well can it be trained to predict the gender distribution of anonymous news commenters for a labeled news article?
- RQ3. How different will the predicted gender distributions for unlabeled news articles be when compared to the true gender distributions for labeled news articles? Particularly, in terms of their histograms and descriptive statistics.

1.4. Organization of This Paper

The rest of this paper is composed of three sections. Section 2 describes the research framework proposed to design and examine a machine-learning-based approach for predicting the gender distribution of anonymous news commenters for a news article. Section 3 demonstrates the results of applying the research framework to the case of the Korean news portal, naver.com. Then, it discusses the performances of different prediction techniques with comparisons. Using the best prediction technique selected, it explores the predicted gender distribution of anonymous news commenters for both labeled and unlabeled news articles. In the end, Section 4 concludes the paper, summarizing the entire study and describing its implications and limitations for future works.

2. Materials and Methods

This paper proposed a machine-learning-based approach to predict the gender distribution of anonymous news commenters for a news article using the textual data of anonymous news comments on the news article. The research framework to design and test our proposed method was suggested, as seen in Figure 1. In addition, the details are explained in the following subsections.

2.1. Data Collection

Data from 1 June to 31 July 2018, were collected from all sections of the Korean news portal, naver.com, by using the developed crawler. The collected data were 14,896,043 news comments for 177,735 news articles, and the textual contents of the 14,896,043 news comments were collected.

In addition, because a news article with more than 100 direct news comments was given gender distribution by naver.com, the gender distributions were collected for such labeled news articles. The male and female rates as the gender distribution for a news article n can be defined by

$$malerate(n) = \frac{n(\mathrm{MALE}(n))}{n(\mathrm{MALE}(n) \cup \mathrm{FEMALE}(n))}, \quad (1)$$

$$femalerate(n) = \frac{n(\mathrm{FEMALE}(n))}{n(\mathrm{MALE}(n) \cup \mathrm{FEMALE}(n))}, \quad (2)$$

where $\mathrm{MALE}(n)$ is a set of male news commenters on n and $\mathrm{FEMALE}(n)$ is a set of female news commenters on n.

As shown in Table 3, it turned out that only 8.7023% of the collected news articles were given the gender distribution and so they were the labeled news articles. In comparison, the rest 91.2977% were not provided with the gender distribution and were the unlabeled news articles.

Table 3. The collected news articles: labeled vs. unlabeled.

Type	Count	Percentage (%)
Labeled	15,467	8.7023
Unlabeled	162,268	91.2977
Total	177,735	100

Figure 2 shows the kernel density estimation (KDE) plot visualizing the distribution of the gender distribution given for the labeled news articles, which is 8.7023% of the collected news articles. Moreover, it was found that 85.9049% of the collected news comments, i.e., 12,796,430 among 14,896,043, belong to the labeled news articles. Table 4 shows the descriptive statistics on the number of news comments per news article respectively for three groups: labeled, unlabeled, and total news articles.

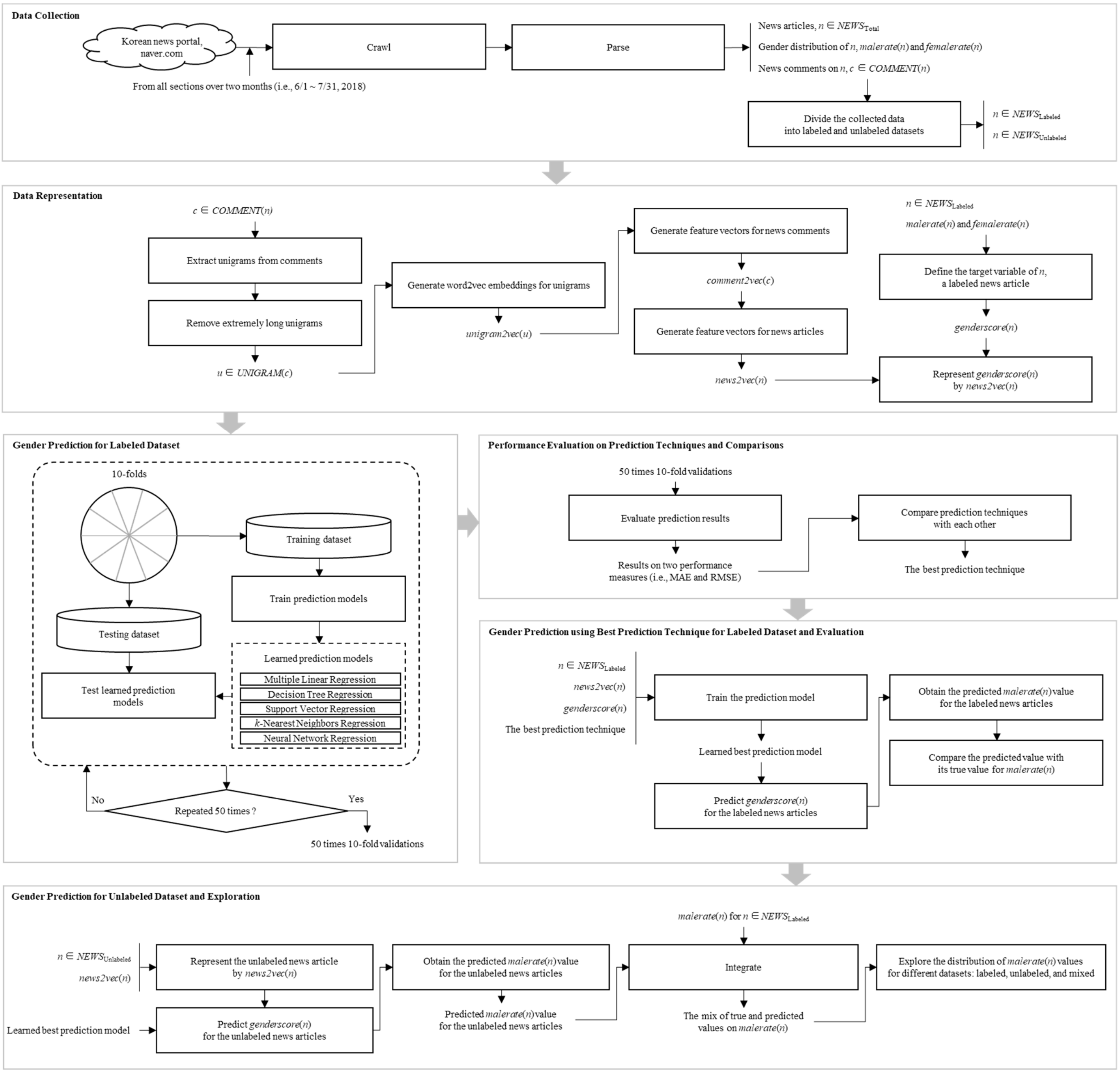
Data Collection
Korean news portal, naver.com
Crawl
Parse
From all sections over two months (i.e., 6/1 ~ 7/31, 2018)
News articles, $n \in NEWS_{\text{Total}}$
Gender distribution of n, $malerate(n)$ and $femalerate(n)$
News comments on n, $c \in COMMENT(n)$
Divide the collected data into labeled and unlabeled datasets
$n \in NEWS_{\text{Labeled}}$
$n \in NEWS_{\text{Unlabeled}}$
Data Representation
$c \in COMMENT(n)$
Extract unigrams from comments
Remove extremely long unigrams
$u \in UNIGRAM(c)$
Generate word2vec embeddings for unigrams
$unigram2vec(u)$
Generate feature vectors for news comments
$comment2vec(c)$
Generate feature vectors for news articles
$news2vec(n)$
$n \in NEWS_{\text{Labeled}}$
$malerate(n)$ and $femalerate(n)$
Define the target variable of n, a labeled news article
$genderscore(n)$
Represent $genderscore(n)$ by $news2vec(n)$
Gender Prediction for Labeled Dataset
10-folds
Training dataset
Train prediction models
Testing dataset
Learned prediction models
Multiple Linear Regression
Decision Tree Regression
Support Vector Regression
k-Nearest Neighbors Regression
Neural Network Regression
Test learned prediction models
No
Repeated 50 times ?
Yes
50 times 10-fold validations
Performance Evaluation on Prediction Techniques and Comparisons
50 times 10-fold validations
Evaluate prediction results
Results on two performance measures (i.e., MAE and RMSE)
Compare prediction techniques with each other
The best prediction technique
Gender Prediction using Best Prediction Technique for Labeled Dataset and Evaluation
$n \in NEWS_{\text{Labeled}}$
$news2vec(n)$
$genderscore(n)$
The best prediction technique
Train the prediction model
Learned best prediction model
Predict $genderscore(n)$ for the labeled news articles
Obtain the predicted $malerate(n)$ value for the labeled news articles
Compare the predicted value with its true value for $malerate(n)$
Gender Prediction for Unlabeled Dataset and Exploration
$n \in NEWS_{\text{Unlabeled}}$
$news2vec(n)$
Learned best prediction model
Represent the unlabeled news article by $news2vec(n)$
Predict $genderscore(n)$ for the unlabeled news articles
Obtain the predicted $malerate(n)$ value for the unlabeled news articles
Predicted $malerate(n)$ value for the unlabeled news articles
$malerate(n)$ for $n \in NEWS_{\text{Labeled}}$
Integrate
The mix of true and predicted values on $malerate(n)$
Explore the distribution of $malerate(n)$ values for different datasets: labeled, unlabeled, and mixed

Figure 1. Research framework, proposed by this study.

Table 4. Descriptive statistics on the number of news comments per news article.

Type	Min	Max	Mean	S.D.
Labeled	100	35,729	827.3376	1417.0208
Unlabeled	1	531	12.9392	23.4438
Total	1	35,729	83.8104	477.4136

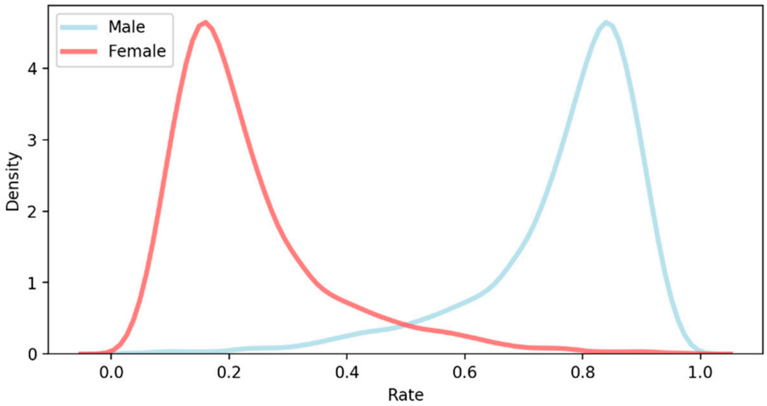

Figure 2. The distributions of *malerate*(*n*) and *femalerate*(*n*) for the labeled news articles.

2.2. Data Representation

As the collected news comments had many irregular texts, applying structural analysis such as POS tagging to the news comments was impossible. Therefore, splitting the text of a news comment on white space, the unigrams of the news comments were extracted as features and used to represent news comments. In other words, using unigrams in a news comment, the news comment was transformed into the BOW form.

Particularly, spacing errors were resolved in the following ways before the tokenization to prevent the extremely long unigrams:

First, the patterns of extremely long unigrams were investigated, and there were found four main types of character entities without spaces as follows: (i) HTML character entities (e.g., & and &gt;); (ii) ending punctuation marks (e.g., ? and !); (iii) characters for text slang such as simple or swearing expressions and emoticons (e.g., ㅎ, ㅋ, and ㅠ); and (iv) unicode characters such as emoji.

Second, for each type, regular expressions to find the extremely long unigrams of the type were generated. Using such regular expressions, the extremely long unigrams that include character entities without spaces were found, and space was inserted before and after the character entities.

After resolving the problem of extremely long unigrams, the feature vector to represent a unigram u was generated for the 16,410,945 unigrams that were identified from the 14,896,043 collected news comments. Here, the feature vectors were obtained by using a word embedding, which is a learned representation for text where words that have the same meaning have a similar representation.

For word embedding, this study adopted word2vec because it is the most popular technique to learn word embeddings using a shallow neural network [26]. In addition, among the two popular algorithms for word2vec, skip-gram and continuous BOW (CBOW), this study selected CBOW as it is faster and has a better representation for more frequent words [27,28]. In addition, to implement word2vec, the word2vec module of GENSIM was used with default settings including 300 as the dimension of word2vec vectors (https://radimrehurek.com/gensim/models/word2vec.html accessed on 15 July 2022).

Thus, the unigrams of a news comment were considered as words in a sentence and input to train the word2vec model. Because the data size of news comments was too big, the 14,896,043 collected news comments were segmented into 100 mini-batches, and the word2vec model was trained incrementally by mini-batch updates. As a result of training the word2vec model, the feature vectors were generated for the generated 16,410,945 unigrams, i.e., *unigram2vec*(u).

Then, to generate text representation for news comments and news articles, this study aggregated the word embeddings of the related unigrams. This approach is preferred for NLP problems as vectorial meaning is its main advantage [7]. In detail, the feature vectors of the 16,410,945 generated unigrams were used to obtain the feature vectors of the 14,896,043 collected news comments, which were subsequently used to generate the feature vectors of the 177,735 collected news articles. Such two types of feature vectors can be defined as follows:

First, *comment2vec*(*c*) is the average of *unigram2vec* vectors from unigrams appearing in a news comment *c* and given by

$$comment2vec(c) = \frac{1}{n(\text{UNIGRAM}(c))} \sum_{u \in \text{UNIGRAM}(c)} unigram2vec(u), \tag{3}$$

where $\text{UNIGRAM}(c)$ is a set of unigrams appearing in c and $u \in \text{UNIGRAM}(c)$.

Second, *news2vec*(*n*) is the average of *comment2vec* vectors from news comments on n and is obtained by

$$news2vec(n) = \frac{1}{n(\text{COMMENT}(n))} \sum_{c \in \text{COMMENT}(n)} comment2vec(c), \tag{4}$$

where $\text{COMMENT}(n)$ is a set of news comments on n.

Moreover, for the 15,467 labeled news articles, the male and female rates of n were put together into a single value, i.e., odds, and its log odds was added as the target variable of this study. It can be defined as

$$genderscore(n) = \log\left(\frac{malerate(n)}{1 - malerate(n)}\right) = \log\left(\frac{malerate(n)}{femalerate(n)}\right). \tag{5}$$

In the end, for a labeled news article, *genderscore*(*n*) as a target variable was represented by using the *news2vec*(*n*) vector as features. In addition, the dataset of 15,467 instances with the values of 300 feature variables and a target variable was constructed as the labeled dataset. On the other hand, the dataset of unlabeled news articles only with 300 feature variables (i.e., without a target variable) was named the unlabeled dataset.

2.3. Gender Prediction for Labeled Dataset

Because the target variable of this study, *genderscore*(*n*), is numerical, machine-learning techniques for prediction, which allow the continuous target variable and were commonly used in previous studies, were selected for this study. Those prediction methods are multiple linear regression (MLR), decision tree regression (DTR), support vector regression (SVR), *k*-nearest neighbors regression (*k*-NNR), and neural network regression (NNR). To implement these five prediction techniques, Python codes were programmed based on the machine learning package scikit-learn, which is a well-known open-source toolkit for machine-learning projects (https://scikit-learn.org accessed on 15 July 2022).

Before conducting experiments, hyperparameters for each regressor were selected, and optimized using the grid search method with 10-fold cross-validation. Then, using the optimized hyperparameters, 10-fold cross-validation for each regressor was performed as an experiment, and the same experiment was repeated 50 times. For each of the experimental repetitions, a different random seed was used, but the random seed was kept identical for the same iteration of different prediction techniques, by referring to the previous studies [2,29,30].

2.4. Performance Evaluation of Prediction Techniques and Comparisons

In this component, the performances of the five prediction techniques were evaluated based on the prediction results. For the performance evaluation, mean absolute error (MAE) and root mean squared error (RMSE) were used as they have been adopted by previous studies whose target variables had continuous values [31]. Smaller values on the performance measures mean less prediction error and indicate better performance [29]. Moreover, to investigate the effect of different prediction techniques on the two performance measures in a statistical way, pairwise *t* tests were performed, and the best prediction technique was selected.

2.5. Gender Prediction Using Best Prediction Technique for Labeled Dataset and Evaluation

In the previous step, the best prediction technique was identified, and in this subsection, a prediction model that uses the best prediction technique was trained from the labeled dataset. Then, the trained prediction model was again applied to predict the $genderscore(n)$ for each news article in the labeled dataset. The predicted $genderscore(n)$ value for the news article in the labeled dataset was transformed into the predicted $malerate(n)$ value, as well as the predicted $femalerate(n)$ value, by referring to Equation (5).

Then, this study evaluated whether the best prediction technique could predict the $genderscore(n)$ for news articles in the labeled dataset accurately or not. To do so, the distribution of the predicted $malerate(n)$ values, obtained from the predicted $genderscore(n)$ values, was compared with the distribution of true $malerate(n)$ values by drawing and investigating histograms. In addition, the paired t-test was used to find out whether the predicted $malerate(n)$ value could be considered equal to the true $malerate(n)$ value for each news article in the labeled dataset in a statistically significant way.

2.6. Gender Prediction for Unlabeled Dataset and Exploration

The learned prediction model, trained by using the best prediction technique and the labeled dataset in the previous step, was applied to predict the $genderscore(n)$ for a news article in the unlabeled dataset. In addition, the predicted $genderscore(n)$ value was transformed into the predicted $malerate(n)$ and $femalerate(n)$ values for the unlabeled news articles. Thus, the unknown gender distribution could be filled up with the predicted $malerate(n)$ and $femalerate(n)$ values for the unlabeled news article. Eventually, the predicted gender distribution of the unlabeled news article could resolve the incomplete gender information problem of the collected news articles.

On the other hand, because there was no true value available for $malerate(n)$ in the unlabeled dataset, the prediction result on $malerate(n)$ for the unlabeled dataset could not be evaluated in terms of MAE and RMSE. Hence, instead of obtaining the MAE and RMSE, the predicted $malerate(n)$ values of the unlabeled news articles were explored by visualization, and their distribution was compared with the distribution of the true $malerate(n)$ values of the labeled news articles. Moreover, it was compared with the mixed $malerate(n)$ values, obtained by integrating the true $malerate(n)$ values of the labeled dataset and the predicted $malerate(n)$ values of the unlabeled dataset. In addition, differences among the three datasets (i.e., labeled, unlabeled, and mixed datasets) were observed in terms of normality (i.e., mean and variance), skewedness, and kurtosis.

3. Results and Discussions

3.1. Evaluation Results of Prediction Techniques

Using the labeled dataset, 10-fold cross-validation as an experiment was repeated 50 times for each prediction technique. The experimental results were evaluated regarding the two performance measures, MAE and RMSE. Then, for the five prediction techniques, the average value of each performance measure over the repeated experiments was obtained, as shown in Table 5. Consequently, Table 5 shows that NNR gave the best results among the five prediction techniques, as it showed a smaller value than the other prediction techniques in terms of both performance measures. This indicates that NNR was best at learning the complex relationship between w2v-based feature vectors and the target variable for the labeled news articles.

3.2. Comparisons between Prediction Techniques

Table 6 shows the comparison results of pairwise t tests, which were performed to evaluate the effects of different prediction techniques. In the hypotheses of Table 6, the greater than sign '>' indicates that the former prediction technique had a bigger prediction error than the latter one when they were compared. In addition, the comparison results can be interpreted as follows: first, if the p-value is statistically significant, the positive value of t means the corresponding hypothesis is supported, but the negative value of t means the opposite of the corresponding hypothesis is supported; and second, if the *p*-value is not statistically significant, the corresponding hypothesis is not supported, and it means there is not sufficient evidence for the inequality.

Table 5. Evaluation results.

Performance Measure	**MLR**		**DTR**		**SVR**		***k*-NNR**		**NNR**	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
MAE	0.0476	0.0004	0.0638	0.0006	0.0476	0.0004	0.0491	0.0004	**0.0416**	0.0005
RMSE	0.0675	0.0007	0.0962	0.0012	0.0675	0.0007	0.0723	0.0009	**0.0579**	0.0008

Note: The best evaluation result is highlighted in bold.

Table 6. Comparison results on the different prediction techniques.

Hypothesis	**MAE**		**RMSE**		**Supported**
	t	***p*-Value**	***t***	***p*-Value**	
MLR > DTR	−166.4182	0.0000 ***	−140.9784	0.0000 ***	Yes
MLR > SVR	0.5370	0.5925	0.0352	0.9720	No
MLR > *k*-NNR	−18.3839	0.0000 ***	−28.3703	0.0000 ***	Yes
MLR > NNR	64.3545	0.0000 ***	63.1935	0.0000 ***	Yes
DTR > SVR	166.9406	0.0000 ***	140.9461	0.0000 ***	Yes
DTR > *k*-NNR	147.7913	0.0000 ***	109.4373	0.0000 ***	Yes
DTR > NNR	205.9550	0.0000 ***	185.0390	0.0000 ***	Yes
SVR > *k*-NNR	−18.9177	0.0000 ***	−28.3842	0.0000 ***	Yes
SVR > NNR	63.9198	0.0000 ***	63.1132	0.0000 ***	Yes
k-NNR > NNR	79.0892	0.0000 ***	83.6033	0.0000 ***	Yes

Note: The significance level is *** $p < 0.01$.

Consequently, at the *p*-value = 0.05, it is seen from Table 6 that there exist statistically significant differences between the prediction techniques except between SVR and MLR in terms of both performance measures. In detail, the five prediction techniques could be arranged as 'DTR > MLR = SVR > *k*-NNR > NNR' in descending order of either MAE or RMSE. Because the smallest value in both performance measures indicates the best prediction technique, NNR was evaluated as the best prediction technique in this study. The comparison results provide statistical evidence for the finding from Table 6, which is that NNR is the best prediction technique. Therefore, NNR was selected to be trained with the labeled dataset for predicting the gender distribution of news commenters for a news article in the unlabeled dataset.

3.3. Results of Gender Prediction for Labeled Dataset and Evaluation

NNR, selected as the best prediction technique, was trained and used to predict the gender distribution of news commenters for a news article in the labeled dataset. As a result, Figure 3 describes the distribution of the $malerate(n)$ values, obtained from the predicted $genderscore(n)$ values by the learned NNR. When compared, for the labeled news articles, the predicted $malerate(n)$ values turned out to have almost the same distribution as the true $malerate(n)$ values.

Moreover, the pairwise t test between the predicted $malerate(n)$ values and their true values for the news articles in the labeled dataset gave $t = -1.3308$ and *p*-value = 0.1833. This means there was no statistically significant difference between them at the level of *p*-value = 0.05.

Thus, these evaluation results indicate that NNR as the best prediction technique could predict the gender distribution of a news article in the labeled dataset in an accurate manner. Moreover, this implies that the gender prediction for the unlabeled dataset can be made successfully and the predicted result for the unlabeled dataset can be used reliably for filling up the unknown gender information in the collected news articles.

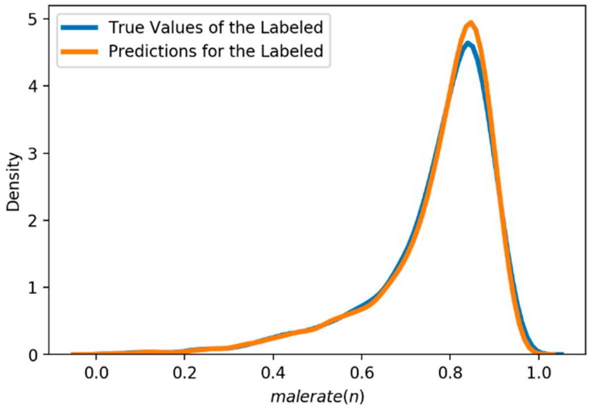

Figure 3. Histogram comparison between the true and predicted values of *malerate*(n) for the labeled news articles.

3.4. Results of Gender Predictions for Unlabeled Dataset and Exploration

The unknown *malerate*(n) values for news articles in the unlabeled dataset were estimated by the prediction model that was learned from the labeled dataset by the selected best prediction technique, NNR. Figure 4 shows the distribution of the predicted *malerate*(n) values for the unlabeled news articles and its comparison with the labeled and mixed datasets. Table 7 shows more detailed comparison results among the three types of *malerate*(n) values.

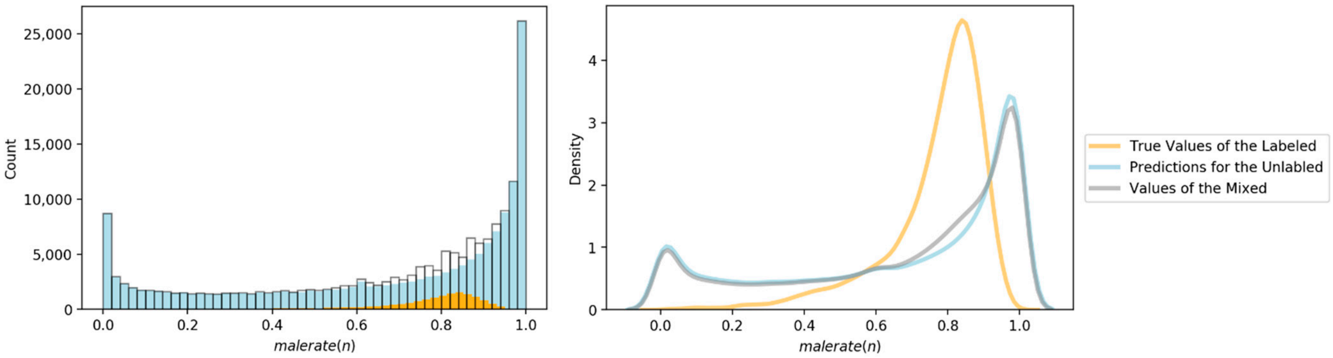

Figure 4. Histograms (**left**) and KDE plots (**right**) on the distribution of the *malerate*(n) for the different datasets.

Table 7. Descriptive statistics on the three types of *malerate*(n) values.

Type	Count	Mean	Variance	Skewedness	Kurtosis
True values of the labeled news articles	15,466	0.7675	0.0198	−1.6953	3.4782
Predictions for the unlabeled news articles	162,268	0.6626	0.1115	−0.7492	−0.8679
Values of the mixed	177,734	0.6717	0.1044	−0.8329	−0.6741

In Figure 4, the true *malerate*(n) values for the labeled news articles have a bell-shaped distribution, while the predicted *malerate*(n) values for the unlabeled news articles have a U-shaped distribution. This finding means that the unknown gender distributions of the unlabeled news articles have a multimodal distribution, which is different from the labeled news articles. When they were mixed, the distribution of the mixed values was more like the predicted values of the unlabeled news articles than the true values of the labeled news articles. This shows that the only gender information from the labeled dataset cannot represent the total dataset.

In addition, Figure 4 shows that the predicted *malerate*(n) values for the unlabeled news articles were more frequent with high or low values, which are respectively equivalent to low or high values in terms of *femalerate*(n). Simply put, they were biased toward either males or females. In addition, this can cause a problem that the anonymous news comments of the unlabeled news articles are considered as public opinions that represent both genders equally, without a doubt or awareness that they may not represent both genders equally.

4. Conclusions

For gender research through social media big data, obtaining gender information is an essential step. However, anonymity and privacy policy have made it difficult or impossible to acquire gender information from social media. To deal with it, focusing on a Korean news portal, naver.com, this paper proposed a machine-learning-based method for predicting the gender distribution of anonymous news commenters for a news article, represented by the w2v-based feature values of the anonymous news comments on the news article.

Using the collected data, the proposed method was evaluated with different prediction techniques. In addition, NNR was selected as the best prediction technique, showing better performances than the others in a statistically significant way and giving an answer to RQ1. Subsequently, the true and predicted *malerate*(n) values for the labeled news articles were compared with each other. By doing so, this study showed the machine-learning-based approach can lead to accurate prediction, solving the incomplete gender information problem of anonymous news commenters, and RQ2 could be resolved. Thus, according to this study, even a dataset without gender information can also be made usable as social media big data for gender research without being ignored as useless data.

Related to RQ3, the unknown *malerate*(n) values were predicted for the unlabeled news articles by using the prediction model, trained with the best prediction technique and the labeled dataset. Their distribution turned out different from the distribution of true *malerate*(n) values for the labeled news articles, i.e., U-shaped vs. bell-shaped. Consequently, this indicates that using only a labeled dataset for gender research can result in misleading findings and distorted conclusions. Therefore, the unlabeled dataset should not be ignored for better and more accurate gender research with social media big data such as news articles and news comments. This has an implication for big data research projects, which have no choice but to use only a small portion of total data.

Overall, the proposed method in this paper can be used as the initial step for future gender research, which uses anonymous social media big data and needs gender prediction to make up for the incomplete gender information. Moreover, it can be extended to many areas related to gender research. For example, gender information predicted for an unlabeled news article from its anonymous news comments can be used to recommend a news article in which the gender distribution of news commenters is not biased toward any gender. In addition, it can be used to evaluate the effectiveness of an advertisement, embedded in a news article to target a specific gender distribution. Moreover, this paper has implications that, even if social media big data, generated from its anonymous users, are incomplete, the incomplete data can be resolved using machine learning. The prediction of the incomplete data can be used to better understand the anonymous social media users as humans and can be helpful practically in keeping societies more sustainable. Thus, this study can contribute to sustainable communities in the future by pioneering a new way for data-driven computational social science with incomplete and anonymous social media big data.

Based on this study, further research can be made, and it will help overcome the limitations of this study and theoretically contribute to gender research for sustainability:

First, the best prediction model, trained by using the anonymous news comments of naver.com, can also be transferred to predict gender information in other anonymous social media. Therefore, whether the proposed approach in this paper can perform well for other anonymous social media needs to be investigated in future work.

Second, this study mainly focused on the Korean language, but its proposed method can also be applied to other languages because it employed a language-independent approach for text data representation, that is, word embeddings. Hence, by further research, the appropriateness of applying this study's proposed method to other languages or multilanguage can be investigated.

Third, the prediction results enabled us to obtain gender information on anonymous news commenters for the unlabeled news articles. This will provide opportunities for further research to analyze and understand gender in social media big data from different perspectives: from words to topics, from individuals to collectives, and from at the moment to over time.

Funding: This study was supported by the National Research Foundation grant (NRF-2021R1F1A1063681) funded by the Korean Government. The APC was funded by Gyeongsang National University.

Data Availability Statement: The data presented in this study are openly available on naver.com. However, the data can not be shared without the permission of naver.com.

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

References

1. Suh, J.H.; Park, C.H.; Jeon, S.H. Applying text and data mining techniques to forecasting the trend of petitions filed to e-People. *Expert Syst. Appl.* **2010**, *37*, 7255–7268. [CrossRef]
2. Suh, J.H. Forecasting the daily outbreak of topic-level political risk from social media using hidden Markov model-based techniques. *Technol. Forecast. Soc. Change* **2015**, *94*, 115–132. [CrossRef]
3. Suh, J.H. SocialTERM-Extractor: Identifying and Predicting Social-Problem-Specific Key Noun Terms from a Large Number of Online News Articles Using Text Mining and Machine Learning Techniques. *Sustainability* **2019**, *11*, 196. [CrossRef]
4. Tsao, S.-F.; Chen, H.; Tisseverasinghe, T.; Yang, Y.; Li, L.; Butt, Z.A. What social media told us in the time of COVID-19: A scoping review. *Lancet Digit. Health* **2021**, *3*, e175–e194. [CrossRef]
5. Bazzaz Abkenar, S.; Haghi Kashani, M.; Mahdipour, E.; Jameii, S.M. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telemat. Inform.* **2021**, *57*, 101517. [CrossRef] [PubMed]
6. Hirt, R.; Kühl, N.; Satzger, G. Cognitive computing for customer profiling: Meta classification for gender prediction. *Electron. Mark.* **2019**, *29*, 93–106. [CrossRef]
7. López-Santillán, R.; Montes-Y-Gómez, M.; González-Gurrola, L.C.; Ramírez-Alonso, G.; Prieto-Ordaz, O. Richer Document Embeddings for Author Profiling tasks based on a heuristic search. *Inf. Process. Manag.* **2020**, *57*, 102227. [CrossRef]
8. Wu, C.; Wu, F.; Qi, T.; Liu, J.; Huang, Y.; Xie, X. Neural Gender Prediction in Microblogging with Emotion-aware User Representation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2401–2404.
9. Reddy, T.R.; Vardhan, B.V.; Reddy, P.V. N-Gram Approach for Gender Prediction. In Proceedings of the 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 5–7 January 2017; pp. 860–865.
10. Kucukyilmaz, T.; Deniz, A.; Kiziloz, H.E. Boosting gender identification using author preference. *Pattern Recognit. Lett.* **2020**, *140*, 245–251. [CrossRef]
11. López-Monroy, A.P.; González, F.A.; Solorio, T. Early author profiling on Twitter using profile features with multi-resolution. *Expert Syst. Appl.* **2020**, *140*, 112909. [CrossRef]
12. Das, S.; Paik, J.H. Context-sensitive gender inference of named entities in text. *Inf. Process. Manag.* **2021**, *58*, 102423. [CrossRef]
13. Cheng, N.; Chandramouli, R.; Subbalakshmi, K.P. Author gender identification from text. *Digit. Investig.* **2011**, *8*, 78–88. [CrossRef]
14. Aman, J.J.C.; Smith-Colin, J.; Zhang, W. Listen to E-scooter riders: Mining rider satisfaction factors from app store reviews. *Transp. Res. Part D Transp. Environ.* **2021**, *95*, 102856. [CrossRef]
15. Lee, S.Y.; Ryu, M.H. Exploring characteristics of online news comments and commenters with machine learning approaches. *Telemat. Inform.* **2019**, *43*, 101249. [CrossRef]
16. Otterbacher, J. Gender, writing and ranking in review forums: A case study of the IMDb. *Knowl. Inf. Syst.* **2013**, *35*, 645–664. [CrossRef]
17. Bamman, D.; Eisenstein, J.; Schnoebelen, T. Gender identity and lexical variation in social media. *J. Socioling.* **2014**, *18*, 135–160. [CrossRef]
18. Choi, Y.; Kim, Y.; Kim, S.; Park, K.; Park, J. An on-device gender prediction method for mobile users using representative wordsets. *Expert Syst. Appl.* **2016**, *64*, 423–433. [CrossRef]
19. Hosseini, M.; Tammimy, Z. Recognizing users gender in social media using linguistic features. *Comput. Hum. Behav.* **2016**, *56*, 192–197. [CrossRef]

20. Teso, E.; Olmedilla, M.; Martínez-Torres, M.R.; Toral, S.L. Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. *Technol. Forecast. Soc. Change* **2018**, *129*, 131–142. [CrossRef]
21. Al-Ghadir, A.I.; Azmi, A.M. A Study of Arabic Social Media Users—Posting Behavior and Author's Gender Prediction. *Cogn. Comput.* **2019**, *11*, 71–86. [CrossRef]
22. Hussein, S.; Farouk, M.; Hemayed, E. Gender identification of egyptian dialect in twitter. *Egypt. Inform. J.* **2019**, *20*, 109–116. [CrossRef]
23. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; 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]
24. Rafique, I.; Hamid, A.; Naseer, S.; Asad, M.; Awais, M.; Yasir, T. Age and Gender Prediction using Deep Convolutional Neural Networks. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–6.
25. García-Díaz, J.A.; Cánovas-García, M.; Colomo-Palacios, R.; Valencia-García, R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.* **2021**, *114*, 506–518. [CrossRef]
26. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
27. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; Volume 2, pp. 3111–3119.
28. Rong, X. Word2vec parameter learning explained. *arXiv* **2014**, arXiv:1411.2738.
29. Choi, B.; Suh, J.H. Forecasting Spare Parts Demand of Military Aircraft: Comparisons of Data Mining Techniques and Managerial Features from the Case of South Korea. *Sustainability* **2020**, *12*, 6045. [CrossRef]
30. Suh, J.H. Comparing writing style feature-based classification methods for estimating user reputations in social media. *SpringerPlus* **2016**, *5*, 261. [CrossRef]
31. Zhang, F.; Gong, T.; Lee, V.E.; Zhao, G.; Rong, C.; Qu, G. Fast algorithms to evaluate collaborative filtering recommender systems. *Knowl. Based Syst.* **2016**, *96*, 96–103. [CrossRef]