Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
15/2 (2022), 131-141. DOI: http://dx.doi.org/10.21609/jiki.v15i2.1079

Gender Prediction of Indonesian Twitter Users Using Tweet and Profile Features

Rahmad Mahendra, Hadi Syah Putra, Douglas Raevan Faisal, Fadzil Rizki
[1] Faculty of Computer Science, Universitas Indonesia, Kampus UI Depok, 16424, Indonesia

Email:rahmad.mahendra@cs.ui.ac.id

Abstract

The increasing use of social media generates huge amounts of data which in turn triggers research into social media analytics. Social media contents can be analyzed to explore public opinion on an issue or provide the insights reflecting proxy indicators towards real-world events. Understanding the demographics of social media users can increase the potential for applications of sentiment analysis, topic modeling, and other analytical tasks. To map demographics, we need to know the latent attributes of users, such as age, gender, occupation and location of residence. Since this attribute is not directly available, we need to do some inference from the social media data. This study aims to predict the gender attribute given a Twitter user account. We conducted experiments with several supervised classifiers with feature extraction, including the use of word embedding representations. The results of this study indicate that the combination of features extracted from Tweet contents and user profile structured data can predict the gender of Twitter users in Indonesia with accuracy above 80%.

Keywords: *gender*, *Twitter*, *user*, *classification*, *feature extraction*, *demography*

1. Introduction

Over the past two decades, social media has gone through rapid development and became an important part of the lives of millions of people around the world. In February 2022, there were 4.62 billion active social media users, which equals 58.4% of the global population. The number of social media users has increased 10.1% compared to the number in the year of 2021 [1]. The interaction between people in social media produces a huge amount of data that can be harnessed to gain knowledge and insight in many aspects of life, such as social, political, and economics.

This research focuses on Twitter, one of the most popular social media that allow users to post tweets, the message which are limited to 280 characters. This type of message makes Twitter a fast-paced information source. Twitter data has been utilized for analytics purpose, such as sentiment analysis [1–3], stance detection [4, 5], topic modeling [6, 7], and information extraction [8–10].

In 2021, there were approximately 500 million tweets sent per day. Even though Twitter has abundant tweets data, the user data is limited and often not explicit. Beside username, there are several fields, like name, bio (short description), and location that can reveal the user demography. However, there is no obligation for the users to disclose all these attributes in their profiles.

Most other user attributes are latent, whereas there are many advantages that can be gained by harnessing those data. Demographic attributes, such as gender and age, are useful for analysis purposes such as recommender system, opinion mining, and market research. A company can analyze the potential sales in certain areas or design a promotion strategy targeting specific demographic groups [11]. The government can harness the demographic attributes to determine certain public policy. In politics, demographic attributes are useful in arranging campaign strategy to win the public vote for particular candidates [12]. Demographic information can be used to examine differential patterns in attitudes and behaviors in social media data [13]. Therefore, the automatic inferring of social media user profile

[1] https://wearesocial.com/uk/blog/2022/01/digital-2022/

attributes becomes an important thing to explore.

Predicting attributes of social media users has been a growing area of research. A number of academic studies have proposed methods for inferring the user demographics (e.g., gender, age, ethnicity, and political affiliation) using information extracted from profile and content posts [14–18]. Most of these studies focused on global user (or big countries, such as US and European countries) as used the tweet in English. For Indonesian context, there have been very limited work on Twitter user latent attributes prediction. This paper presents our work on the task of predicting Indonesian Twitter users' gender.

2. Related Work

Rao et.al. [14] and Zamal et.al. [16] used Support Vector Machine classifier to infer age, gender, and political view of Twitter users. Rao et.al. utilized sociolinguistic feature groups and n-grams. The sociolinguistic features include, but not limited to, emoticons, repetitive alphabet, capitalization, and exclamation. While the n-gram features are extracted from the vocabulary in the user's tweet [14]. While, Zamal et.al. [16] employed the number of features for prediction, such as words, n-grams, out/in-neighborhood size, hashtags, mentions, and retweet frequencies. Pennacchiotti & Popescu [19] observed the user behavior, network structure and the linguistic content of the user's Twitter feed to identify political affiliation, ethnicity, and business affirnity.

Burger et.al. [15] worked on a gender classification task on Twitter users. They extracted n-gram features from user tweets and three text fields in Twitter user profile (screen name, full name, and short description). They experimented with Naive Bayes, Support Vector Machine, and Balanced Winnow2. Alowibdi et.al. [20, 21] proposed color-based features to detect gender of Twitter user. In addition, they also proposed phoneme technique to extract features from user name. Liu and Ruth [17] studied the use of first name as the features for gender classification.

For Indonesian Twitter users, there exists few previous works [22–24]. Wibisono & Faruqi [22] built a gender classification classifier by using sociolinguistic and lexical features. While, Rasis et.al. [23] employed Naive Bayes model to infer gender of Indonesian Twitter user using the features consisting of screen name, full name, and tweets.

Siswanto & Khodra [25] worked on predicting the latent attributes of Indonesian Twitter users by using lexical features that were built from the users' tweets. The predicted attributes were age (below 20 or above 20) and job (student or employee). This research utilized data from 300 users with a maximum of 3200 tweets from each user. Experiments were carried out by using three classifiers, namely SVM, Naive Bayes, and Random Forest. Hawari & Khodra [24] conducted a research to predict gender, age, and interest by extracting lexical and sociolinguistic features from tweets. They found that the model with lexical features outperforms the model with sociolinguistics features.

Our study extends previous work on inferring gender label for Indonesian Twitter user. We propose richer hand-crafted features for classification and explore the word embedding representation for lexical features. We run experiment in larger dataset of more than 10K Indonesian Twitter user accounts.

3. Methods

We formulate the task of gender prediction of Indonesian Twitter account, in which we want to predict whether the owner of a user account is male or female, as a binary classification problem, Formally, given a set of Twitter user accounts $U = \{U_1, U_2, ..., U_n\}$ and a set of label $L = \{male, female\}$, we seek a classifier function $f : U \rightarrow L$. We apply supervised learning approach. Therefore, devising hand-crafted features is one of the most important steps in our task.

3.1. Classifiers

Our gender prediction is built on eight different supervised classifiers, namely Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine [26], AdaBoost [27], Gradient Boosting [28], and Multilayer Perceptron. All classifiers are implemented using Python ScikitLearn[2]

3.2. Features

We extract the hand-crafted features from profile and tweets data. In brief, we incorporate eight categories of features: name, user name, user description, color, social network, tweet behavior and interaction, tweet sociolinguistics, and tweet text features.

1. *Name features* (NAME)

While several names are gender-neutral, e.g. 'Eka' and 'Ade', most Indonesian names may reflect someone's gender. For instance, 'Hendra', 'Putra', and 'Muhammad' are male, whereas 'Ayu' and 'Yanti' are female.

[2]https://scikit-learn.org/

The Twitter user profile has a name attribute which is usually filled with first and last name of the user. Although the Twitter user may fill in any string for name (no real name), Peddinti et.al. [29] observed that nearly 70% name in Twitter can still be identifiable as person name. To validate the name, we apply naive name detection algorithm. We inspect whether the name field consists of non alphabet characters (e.g., number or punctuation). We tokenize the name and check whether any token is stopword.

We leverage the value of name field as two features under NAME category, i.e. bag-of-name and name dictionary.

(a). *Bag-of-Name* features

We propose three different representations of bag-of-name features. For first representation, we suppose name as a natural language sentence. We tokenize the name into bag of words. Instead of full name, we only consider first name as the feature in second representation. In third representation, we parse name into first name, middle name, and last name. If the name consists of more than 3 words, middle name has more than one feature value. If the name is single word, both middle and last names are treated as null values. The feature size for first and second representation is the number of vocabulary, while the feature size is three times larger for third representation.

(b). *Name Dictionary* features

We compile dictionary by harvesting 23,366 male and 12,134 female names from Wikidata. Dictionary features for gender prediction experiment correspond to bag-of-name features. Dictionary feature values indicate whether bag-of-name entries found in name dictionary.

2. *Username features* (USERNAME)

Username is a unique identifier of a Twitter user's account. It starts with "@" symbol. To construct the features from username, we propose two different representations, i.e., unigram and char-n-gram. First, we pre-process the username if it contains non alphabet characters. For example, removing such numbers that appears as suffix in the username (e.g., "*putra123*" $\rightarrow$ "*putra*"). Pre-processed username becomes unigram features. For character-level n-gram, we search substring space of username whose length is n. In this research, we assign the value of n as $\{3, 4, 5\}$. To illustrate, the unigram "*putra*" produces several char-n-grams: "*put*", "*utr*", "*tra*", "*putr*", "*utra*", and "*putra*".

3. *User description features* (BIO)

Bio is a short description that is displayed on the user profile page. It should be not more than 160 characters length. Bio field usually describes the character, preference, or personality of the user. We evaluate five different representations for BIO features label=()

1) bag-of-words (BoW)
2) bag-of-words with stopword removal (BoW *stopw*)
3) Word2Vec embedding pre-trained from [30][3]
4) FastText embedding pre-trained from emot data [30]
5) FastText embedding pre-trained from Indo4B data [31]

The size of BoW features is limited not to exceed 100,000. The vector sizes of Word2Vec, emot FastText, and FastText-Indo4B used in the experiments are 400, 100, and 300, respectively.

4. *Color features* (COLOR)

Twitter allows the users to personalize their profile, including color scheme in the profile. The users can choose the color of the page background, link text, sidebar border, sidebar fill, and text. Alowibdi et.al. [20] found that the profile color preference has a fairly high correlation with the user gender. We utilize five aforementioned types of color as the COLOR features in this research.

The profile colors in Twitter are stored in the form of six hexadecimal digits representing 256^3 RGB (red, green, blue) combinations. Due to the large amount of color combinations, color reduction method was applied to reduce the color size from 256 bit to 8 bit. Thus, each COLOR feature only has the size of 512 (8^3).

5. *Social Network features* (NETWORK)

Twitter characteristic as a social network can be utilized as the feature to infer user latent attribute. To understand the social network features, we provide the following Twitter terminologies.

- Follow activity in Twitter is doing a subscription to other users' tweets. If the user A is a follower of the user B, the user A can see user B's tweets on the Twitter timeline.
- Retweet activity is reposting other users' tweet.
- Twitter Lists allow a user to customize, organize and prioritize the Tweets they want to see in the timeline. A user can choose to join

[3]https://github.com/meisaputri21/Indonesian-Twitter-Emotion-Dataset

Lists created by others on Twitter. A user can also create Lists of other accounts by group, topic or interest.

We employ six NETWORK features, i.e., number of followers, number of followees, the listed count, favorites count, average number of retweets (how many tweets being retweeted by other user per number of tweets), and average number of likes (how many tweets being liked by other user per number of tweet).

6. *Tweet Behavior and Interaction* (BEHAVIOR)

Several features for attribute prediction task can be derived from the behavior of the Twitter users, especially when interacting with other users or using topic words in their tweets.

label=()

1) number of tweets contain mentions
2) number of tweets with hashtags
3) number of tweets has image
4) number of tweets has video
5) number of retweet messages (how many tweets that are retweet from others)
6) average number of daily tweets

Mention is an act of engaging other users in a tweet, by referring to their username. Mention is marked by the character "@" at the beginning of a mentioned username. Hashtag is a text starting with "#" used to emphasize the topic or important term in a tweet. Hashtag is marked by the character "#" followed by the emphasized word.

7. *Sociolinguistics features* (SOCIO)

The writing style when posting the tweet may reveal the user gender. For example, the women tend to be more emotional than men in online platform, so they usually send more messages with emoticons [32, 33]. Several works have also examined the association between linguistics use and gender difference in social media.

We utilize 11 SOCIO features. Most of them are inspired from features proposed in previous works, [14, 15, 24]. In addition, we also extract the features considering Part-of-Speech information. label=()

1) Word count
2) Character count
3) Exclamation marks
4) Repeated alphabet: words containing three or more repetitive letters in order. For example: "*aaa*", "*hellooo*", "*sebellll*".
5) Ellipses: sequence of three or more periods used to indicate a pause or an unfinished thought.
6) Capitalized words: words started with a capital letter.
7) Uppercase words: words in which each letter is capitalized.
8) Emoticon: facial expression that can enrich messages with emotional context. An emoticon is formed by combination of the keyboard characters, such as frown ":(" or smile ";-)".
9) Adjective words
10) Noun words
11) Verb words

The value for each aforementioned feature is the average number of certain linguistics phenomenon per tweet. For example, to obtain emoticon feature, we count the number of emoticons in all user's tweets and divided by the total number of tweets posted by the user.

8. *Tweet features* (TWEET)

Tweet is a short message published by users that can be seen by other users. At first, tweet was limited to 140 characters. Since November 2017, it has been extended to 280 characters. All user's tweet are concatenated as single document and this document is represented as text features. Variation of TWEET features representation is similar to BIO, i.e., bag-of-words and word embeddings.

4. Experiments

4.1. Data

Data for experiment is obtained from three sources. First, we sample a number of Twitter user accounts from Arafat's work [34]. A human annotator check whether each Twitter handle is real human, and filter out the buzzer account [35].

We also perform data collection. We gather the Twitter user accounts of Indonesian public figures (e.g. celebrities, politicians, and influencers) that are not part of training data. We also collect a subset of public users from our (paper's authors) Twitter network. To construct gold-label data, we avoid guessing the gender merely based on user profile. So, we include only the users that are real-world friends.

Moreover, we collect additional data set of Twitter users from participants in online surveys taken through Google Form. We restrict participation in our surveys to active users based in Indonesia. The participants are asked to provide their public Twitter user accounts and take a demographic questionnaire. The questionnaire does not only ask gender information of user, but also other demographic attributes,

i.e. age (birth month and year), location (origin and residence), ethnicity, mother language, marital status, education, and occupation. The questionnaire is enclosed with consent form regarding data privacy. The survey participants are given the freedom to share the demographic attributes they are willing to share. We hope that other attributes collected in this survey can be used for further research.

In total, we accumulate 12,020 Twitter users, consisting of 6,010 male and female respectively. We split the data into 8,214 instances in training and 3,806 in testing set. We keep balanced gender distribution in both training and testing set.

For each user, we collect the profile and tweets data. Given the Twitter username, we crawl user profile in structured format using tweepy [4] tool. A profile provides customized personal details of a user, including a unique alphanumeric ID identifying the account; a name field which usually contains the user's first and last name; a profile picture; and a URL. On the other hands, a collection of tweets for targeted user is obtained using twint [5] tool. Tweet data is not only Twitter-text, but also completed with metadata, i.e. user who posted the tweet and descriptive statistics (e.g. number of times the tweet has been retweeted and replied to).

4.2. Experimental Setting

We perform several experiments to compare the performance of classifiers with features variation. We investigate the best representation for the features group extracted from tweet and profile description written in free-text. We compare the model when using different text representation and pre-processing steps for `NAME`, `USERNAME`, `BIO`, and `TWEET` features. Furthermore, we conduct the further experiment on `TWEET` features by evaluating the effect of tweet size on the model performance. We sample the tweets for each user and compare the performance with the model using all tweets for user classification.

We perform the ablation study on the non-text features groups extracted from user profile, i.e., `COLOR`, `NETWORK`, `BEHAVIOR`, and `SOCIO`. We apply leave-one out (LOO) technique for feature ablation study. It is done by systematically removing the feature one by one. For example, the `BEHAVIOR` feature group consists of six features. Thus, we build six different one-out-feature combination in ablation study (e.g., `BEHAVIOR`$-\{$`mention`$\}$, `BEHAVIOR`$-\{$`hashtag`$\}$, ...). We compare the performance of each one-out-feature combination with complete feature combination. When the performance of certain one-out-feature combination is worse than complete one, it indicates that the removed feature in one-out-feature combination contribute positively to improve model performance.

Finally, we evaluate the combination of all proposed features by performing the ablation study. We also analyze the trade-off in using advanced features compared to simple ones with regard to training/inference time.

As the data used in the experiment contains balanced number of male and female users, the performance of our proposed model is measured using accuracy metric. Accuracy is number of users whose gender label is correctly predicted by the model divided by the total number of users evaluated.

4.3. Results

Tables 1 to 3 present the experimental result for text features extracted from the user profiles, while Table 4 reports the result for tweet features. For `NAME` features, dictionary is able to improve the model performance in nearly all models. All classifiers have comparable performance for best model (range of 72 – 77%). The best model is Gradient Boosting clasifier using combination bag-of-words derived from full name and dictionary that achieves an accuracy of 77.10%. On the other hand, the character n-gram outperforms unigram representation for `USERNAME` features. Naive Bayes and Logistic Regression classifier increase the accuracy up to more than 15% when `USERNAME` features are represented as char-$\{3, 4, 5\}$ gram. The rationale is that many unique names that share same substring (e.g., "*andin*, *andini*, *dina*" are female name).

The experiment on `BIO` features shows that no significance difference of using FastText compared to traditional bag-of-words. Word2Vec is the worst representation for `BIO` features. On the other hand, the bag-of-words outperform the word embedding with margin of more than 4% in all classifiers using `TWEET` features, except for Random Forest and Naive Bayes.

Figure 1 depicts the result of ablation study on `NETWORK` and `BEHAVIOR` features. Number of followers and followees fail to help the model when the data is trained using Multinomial Naive Bayes, Logistic Regression, or Support Vector Machine (shown by the accuracy improvement when this feature is ablated), while these two features contribute positively in tree classifier family. Number of mention and hashtag features can be excluded

[4] https://www.tweepy.org/

[5] https://github.com/twintproject/twint

TABLE 1
EXPERIMENTAL RESULT FOR NAME FEATURES

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
full name	65.71%	66.44%	57.09%	60.09%	62.25%	68.11%	52.60%	66.54%
+ dict	70.24%	71.40%	76.00%	77.10%	72.84%	75.30%	68.24%	62.95%
first name	62.72%	63.12%	56.46%	57.82%	57.79%	64.28%	51.93%	63.25%
+ dict	73.20%	73.74%	71.60%	74.37%	70.64%	72.87%	74.70%	72.60%
first, middle, last name	64.35%	64.95%	55.39%	58.02%	60.45%	66.28%	51.86%	65.61%
+ dict	65.31%	66.44%	74.80%	74.10%	72.37%	75.13%	67.61%	66.25%

TABLE 2
EXPERIMENTAL RESULT FOR USERNAME FEATURES

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
bag-of-words	50.20%	50.13%	50.03%	50.03%	50.00%	50.17%	50.03%	50.20%
char 3-5 gram	61.45%	63.48%	59.59%	61.88%	69.97%	66.61%	54.16%	62.38%

TABLE 3
EXPERIMENTAL RESULT FOR BIO FEATURES

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
BoW	52,53%	52,03%	51,76%	52,43%	51,96%	53,30%	49,47%	53,26%
BoW + stopwrm	51,66%	52,53%	52,06%	53,03%	52,50%	53,36%	49,80%	53,16%
Word2Vec	50,90%	49,90%	48,80%	48,70%	47,44%	49,83%	48,30%	50,43%
FastText (emot)	50,70%	51,17%	53,70%	53,33%	49,07%	53,00%	49,40%	52,50%
FastText-Indo4B	51,30%	51,13%	53,63%	54,13%	49,10%	53,99%	48,90%	56,26%

TABLE 4
EXPERIMENTAL RESULT FOR TWEET FEATURES

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
BoW	63.88%	60.39%	72.74%	75.77%	48.87%	75.83%	59.99%	76.90%
BoW + stopwrm	64.61%	57.96%	73.37%	76.30%	48.70%	77.30%	59.89%	77.16%
Word2Vec	55.86%	59.02%	61.09%	64.88%	47.17%	70.94%	46.14%	73.80%
FastText (emot)	56.46%	62.35%	66.34%	70.67%	46.04%	63.02%	42.14%	77.20%
FastText-Indo4B	59.22%	63.91%	69.51%	72.67%	46.67%	71.77%	41.44%	71.57%

when building the gender predictor model. Ablating mention feature in Gradient Boosting, AdaBoost, and Naive Bayes classifiers increases the accuracy more than 2%. In other classifiers, whether mention feature is incorporated in the model does not make significant difference statistically. Number of hashtag is also found as not important feature in tree-based classifier.

We find that 10 of 11 proposed SOCIO features contribute positively in Decision Tree and Gradient Boosting. Verb word features are not discriminative to categorize the gender of Indonesian Twitter user and the model performance decreases when the features are used in 7 of 8 classifiers. Among the strong non-text features in all classifiers are average number of tweets posted every day (NETWORK feature), the number of capitalized words and repeated alphabets (SOCIO features), and background profile color (COLOR feature).

Table 5 presents the model performance for each feature group and combination of all features. The model trained using text features can better predict the gender label compared to the model using non-text features. 12 out of 16 models using either NAME or TWEET features achieve accuracy score more than 70%. Most of models using either NETWORK or BEHAVIOR are only able to achieve accuracy slightly above the random baseline.

Among eight classifiers evaluated in this study, the ensemble learners, i.e., Gradient Boosting and AdaBoost, are the top two with finest accuracy in the experiment with different features. The best accuracy for those two are 82.36% and 78.60% when all proposed feature categories are combined. However, the feature combination does not work well for other six classifiers. The results obtained when using all features categories are still below when using only NAME features. While Support Vector Machine often

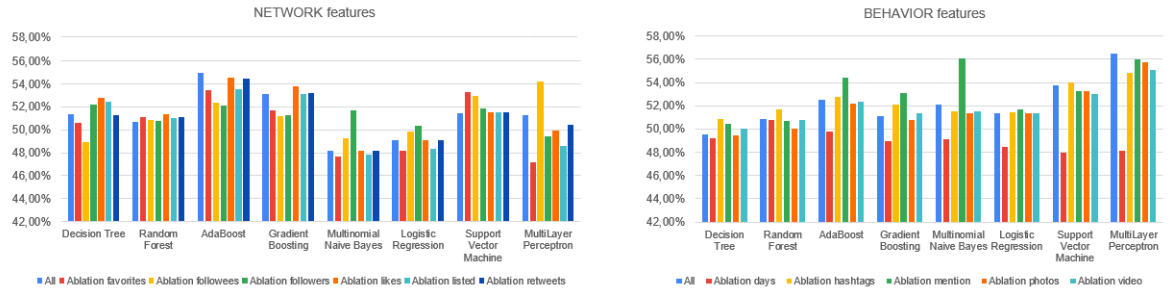

Figure 1. ABLATION STUDY ON NETWORK AND BEHAVIOR FEATURES

TABLE 5
EXPERIMENTAL RESULT FOR ALL FEATURE GROUPS

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
NAME	73.20%	73.74%	76.00%	77.10%	72.84%	75.30%	74.70%	72.60%
USERNAME	61.45%	63.48%	59.59%	61.88%	69.97%	66.61%	54.16%	62.38%
BIO	52,53%	52,56%	53,70%	54,13%	52,50%	53,99%	49,80%	56,26%
COLOR	59.42%	59.55%	62.68%	62.35%	62.48%	62.62%	60.99%	62.15%
NETWORK	52.83%	51.33%	54.99%	53.79%	51.70%	50.40%	53.26%	54.26%
BEHAVIOR	50.83%	51.73%	54.39%	53.13%	56.09%	51.66%	54.03%	56.46%
SOCIO	54.76%	56.96%	60.85%	61.38%	52.03%	56.62%	54.39%	58.56%
TWEET	64.61%	63.91%	73.37%	76.30%	48.87%	77.30%	59.99%	77.20%
all features	68.84%	60.59%	78.60%	82.36%	49.60%	74.53%	58.42%	63.81%

perform the best in the number of previous studies, we have different finding in this study. Except for using NAME features, Support Vector Machine only capable of achieving 60% accuracy or less.

We investigate whether the feature ablation can achieve better performance when combining all features. Figure 2 shows the model accuracy if individual feature group is removed. We find that ablating NETWORK or BEHAVIOR feature upgrade the model performance in 5 classifiers. Logistic Regression achieves accuracy of 79.99% when combining all features but NETWORK. Removing NETWORK also boosts the performance of Multilayer Perceptron and Support Vector Machine up to 7% higher.

We further group the features into three categories, i.e., profile text (combination of NAME, USERNAME, and BIO), profile non-text (combination of COLOR, NETWORK, BEHAVIOR, and SOCIO), and TWEET. For first two feature categories, we apply the number of ablation studies. The best combination for non-text feature group is when the NETWORK feature is not included in 6 classifiers. This finding supports the result of the ablation study reported in Figure 2.

Table 6 presents the experimental result for grouped features. We compare the result with the prediction using combination of all features with ablation study (best result from Figure 2) and NAME feature, the best individual feature observed in Table 5. We find that all classifiers can achieve the accuracy not less than 70% using best feature configuration. Multinomial Naive Bayes performs the best when using text features from user profile, while the best result for Decision Tree is obtained when utilizing all features derived from user profile. Random Forest and Support Vector Machine attain the best accuracy by relying on NAME feature only. Four other classifiers that are top performers in our experiment work the best by employing TWEET features or incorporating them with the features extracted from the profile.

TWEET feature is potential to infer gender label. We use all user tweets so far in current experiment. However, processing all tweets might be computationally expensive. We are interested in investigating how the use of fewer tweets affects the model performance. We rerun the experiments for TWEET feature by using different number of tweets. The number of tweets used as the features does not exceed n ($n = 1, 5, 20, 50, 100, 200, 500, 1000, 2000$). These experiments are carried out on three algorithms dan the result is reported in Table 7.

As expected, the more tweets used, the higher the model's performance. An interesting finding is that FastText excel bag-of-words when using few number of tweets. Moreover, stopword removal makes the model performance trained on bag-of-word TWEET features worse. Those patterns are the opposite when

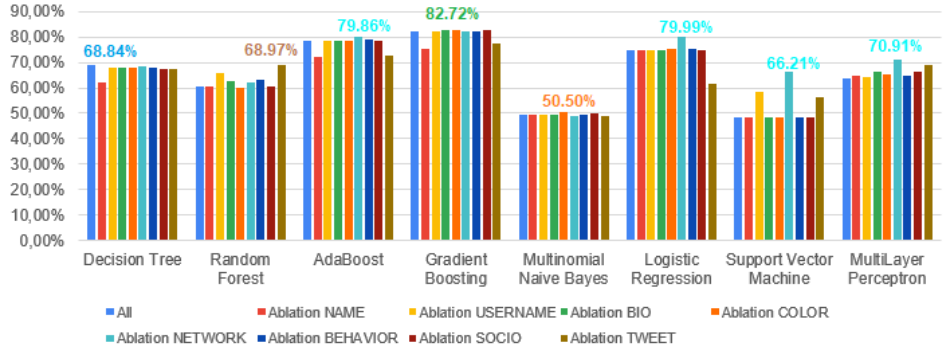

Figure 2. ABLATION STUDY ON COMBINATION OF ALL FEATURES

TABLE 6
EXPERIMENTAL RESULT FOR PROFILE TEXT, PROFILE NON-TEXT, AND TWEET FEATURES

	DT	RF	AdaBo	GBo	NB	LR	SVM	MLP
Profile text	67.14%	71.67%	75.10%	76.96%	76.83%	73.74%	73.17%	69.34%
Profile non-text	56.96%	59.69%	63.45%	63.45%	56.76%	62.75%	54.89%	55.96%
Tweet	64.61%	63.91%	73.37%	76.30%	48.87%	77.30%	59.99%	77.20%
Profile text + non-text	70.57%	69.67%	74.93%	76.40%	71.90%	74.73%	58.89%	66.28%
Profile text + tweet	69.71%	60.55%	79.39%	81.89%	49.60%	78.53%	59.75%	73.83%
Profile non-text + tweet	65.38%	63.15%	73.74%	76.13%	49.13%	77.10%	58.99%	60.72%
All (ablated)	68,84%	68,97%	79,86%	82,72%	50,50%	79,99%	66.21%	70,91%
NAME	73.20%	73.74%	76.00%	77.10%	72.84%	75.30%	74.70%	72.60%

TABLE 7
EXPERIMENT WITH THE TWEETS IN DIFFERENT NUMBER

# tweets	GBo			LR			MLP		
	BoW	BoW + stopwrm	FastText	BoW	BoW + stopwrm	FastText	BoW	BoW + stopwrm	FastText
1	52.13%	50.57%	50.43%	53.13%	52.80%	53.66%	51.27%	51.76%	52.76%
5	52.40%	52.23%	57.96%	56.59%	55.43%	63.22%	57.52%	56.23%	60.59%
10	57.86%	55.56%	60.09%	58.56%	58.29%	65.48%	57.32%	56.69%	62.25%
20	60.39%	57.86%	62.78%	61.48%	60.82%	67.81%	62.38%	62.02%	66.15%
50	64.98%	62.55%	63.91%	65.38%	64.88%	69.24%	66.68%	64.81%	72.84%
100	68.24%	66.94%	66.05%	67.71%	66.68%	70.54%	70.97%	68.24%	72.34%
200	70.84%	70.74%	67.81%	69.11%	68.34%	70.74%	73.27%	72.24%	71.57%
500	73.80%	73.10%	69.84%	68.97%	68.71%	71.04%	73.14%	75.43%	70.64%
1000	74.87%	75.33%	70.57%	75.83%	75.80%	70.07%	77.53%	75.07%	70.17%
2000	76.10%	75.70%	70.47%	75.63%	76.56%	71.57%	75.97%	77.46%	72.54%
all tweets	75.77%	76.30%	72.67%	75.83%	77.30%	71.77%	76.90%	77.16%	71.57%

we all user tweets. On the other hand, using 200 tweets is still able to produce a competitive model, only 5% accuracy point adrift of models that use 10 times more tweets or more. Even a model with TWEET feature that only uses 5 or 10 tweets can perfom better than a model that uses non-text features, such as NETWORK or BEHAVIOR.

We evaluate the computational time to build the model in addition to perfomance accuracy. We observe the time spent to train the model and the time required by the model to process new data and make a prediction. In term of both training and inference time, Support Vector Machine is the most expensive classifier. Gradient Boosting and Multilayer Perceptron are time consuming in the training process, while AdaBoost needs a long time for the inference process. Conversely, the most time-saving classifier is Multinomial Naive Bayes. In terms of features, the usage of non-text features requires less training and inference time. Among the text features,

USERNAME requires the longest time for training.

The top performer in our gender prediction task is Gradient Boosting classifier using all features except BIO (see Figure 2 and Table 6). The time needed for training is more than 15,000 seconds (∼4 hours) and the time for predicting the label for new data is 3 seconds. Other classifier that has competitive performance but with less computational time is Logistic Regression using all features excluding NETWORK. Although the accuracy is 3% lower compared to the Gradient Boosting, the Logistic Regression only require 360 seconds (∼40 times faster) and 1.5 seconds for training and inference steps respectively.

5. Conclusion and Future Work

In this study, we work on gender prediction task on Indonesian Twitter data. We propose the number of features extracted from the information in user profile as well as tweets, i.e. name, username, user bio, profile color, social network, tweet behavior, sociolinguistics, and tweet contents. Although most features contribute positively to predict gender label better than random baseline, combining all features does not always produce the model with better performance. All classifiers can predict with the accuracy more than 70% with appropriate choice of feature ablation. Gradient Boosting classifier employing text features, especially NAME or TWEET, perform the best. Logistic Regression can be considered as an alternative classifier when we concern about computational time. We release the model at https://github.com/ir-nlp-csui/indotwittergender for accelerating the research in application of social media analytics.

We leave the model explainability and thorough error analysis for future work. Pre-trained language model, e.g., IndoBERT [31], can be utilized to improve model performance when using tweet or other text features. On the other hand, multimodal features can be leveraged for gender prediction task as observed in [36, 37]. Predicting other latent attributes can be other research direction to analyze user demography using social media analytics.

Ethical Consideration

Information about gender in general can be regarded as private matters, hence predicting the gender of someone could be regarded as privacy breaching. We encourage that the gender prediction task is not intended to reveal the gender attribute at individual level. On the other hand, this research only analyzes the public user accounts.

References

[1] N. A. Vidya, M. I. Fanany, and I. Budi, "Twitter sentiment to analyze net brand reputation of mobile phone providers," *Procedia Computer Science*, vol. 72, pp. 519–526, 2015.

[2] W. Budiharto and M. Meiliana, "Prediction and analysis of indonesia presidential election from twitter using sentiment analysis," *Journal of Big data*, vol. 5, no. 1, pp. 1–10, 2018.

[3] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 2, pp. 1–29, 2018.

[4] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised user stance detection on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 141–152.

[5] A. H. Nababan, R. Mahendra, and I. Budi, "Twitter stance detection towards job creation bill," *Procedia Computer Science*, vol. 197, pp. 76–81, 2022.

[6] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar, and R. Pranata, "Twitter topic modeling on football news," in *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2018, pp. 467–471.

[7] C. P. S. Kaunang, F. Amastini, and R. Mahendra, "Analyzing stance and topic of e-cigarette conversations on twitter: Case study in indonesia," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 0304–0310.

[8] D. Anggareska and A. Purwarianti, "Information extraction of public complaints on twitter text for bandung government," in *2014 International Conference on Data and Software Engineering (ICODSE)*. IEEE, 2014, pp. 1–6.

[9] P. K. Putra, D. B. Sencaki, G. P. Dinanta, F. Alhasanah, and R. Ramadhan, "Flood monitoring with information extraction approach from social media data," in *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*. IEEE, 2020, pp. 113–119.

[10] P. K. Putra, R. Mahendra, and I. Budi, "Traffic and road conditions monitoring system using extracted information from twitter," *Journal of Big Data*, vol. 9, no. 65, 2022.

[11] T. Hennig-Thurau, C. Wiertz, and F. Feldhaus, "Does twitter matter? the impact of microblogging word of mouth on consumers' adoption

of new movies," *Journal of the Academy of Marketing Science*, vol. 43, no. 3, pp. 375–394, 2015.

[12] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, 2011, pp. 89–96.

[13] T. H. McCormick, H. Lee, N. Cesare, A. Shojaie, and E. S. Spiro, "Using twitter for demographic and social science research: Tools for data collection and processing," *Sociological Methods & Research*, vol. 46, no. 3, pp. 390–421, 2017, pMID: 29033471. [Online]. Available: https://doi.org/10.1177/0049124115605339

[14] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, ser. SMUC '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 37–44. [Online]. Available: https://doi.org/10.1145/1871985.1871993

[15] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1301–1309. [Online]. Available: https://aclanthology.org/D11-1120

[16] F. A. Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, no. 1, 2012, pp. 387–390.

[17] W. Liu and D. Ruths, "What's in a name? using first names as features for gender inference in twitter," in *2013 AAAI Spring Symposium Series*, 2013.

[18] M. Vicente, F. Batista, and J. P. Carvalho, "Twitter gender classification using user unstructured information," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2015, pp. 1–7.

[19] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," in *Proceedings of the international AAAI conference on web and social media*, vol. 5, no. 1, 2011.

[20] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 739–743. [Online]. Available: https://doi.org/10.1145/2492517.2492632

[21] ——, "Empirical evaluation of profile characteristics for gender classification on twitter," in *2013 12th International Conference on Machine Learning and Applications*, vol. 1. IEEE, 2013, pp. 365–369.

[22] Y. Wibisono and N. Faruqi, "Penentuan gender otomatis berdasarkan isi microblog memanfaatkan fitur sosiolinguistik," *Jurnal Cybermatika*, vol. 1, no. 1, 2013.

[23] H. Rasis, A. Erwin, J. Purnama, and M. Galinium, "Automatic demographic classification of indonesian twitter users," in *Proceedings of the 2013 Information Systems International Conference (ISICO)*, December 2013, p. 385–389.

[24] M. A. A. Hawari and M. L. Khodra, "Predicting latent attributes by extracting lexical and sociolinguistics features from user tweets," in *2014 International Conference on Data and Software Engineering (ICODSE)*, 2014, pp. 1–5.

[25] E. Siswanto and M. L. Khodra, "Predicting latent attributes of twitter user by employing lexical features," in *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2013, pp. 176–180.

[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: Association for Computing Machinery, 1992, p. 144–152. [Online]. Available: https://doi.org/10.1145/130385.130401

[27] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[29] S. Peddinti, K. Ross, and J. Cappos, ""on the internet, nobody knows you're a dog": A twitter case study of anonymity in social networks," in *COSN 2014 - Proceedings of the 2014 ACM Conference on Online Social Networks*. Association for Computing Machinery, Inc, Oct. 2014, pp. 83–93.

[30] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion classification on Indonesian Twitter dataset," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 90–95.

[31] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.85

[32] A. Wolf, "Emotional expression online: Gender differences in emoticon use," *Cyberpsychology & behavior*, vol. 3, no. 5, pp. 827–833, 2000.

[33] C. C. Tossell, P. Kortum, C. Shepard, L. H. Barg-Walkow, A. Rahmati, and L. Zhong, "A longitudinal study of emoticon use in text messaging from smartphones," *Computers in Human Behavior*, vol. 28, no. 2, pp. 659–663, 2012.

[34] T. A. Arafat, I. Budi, R. Mahendra, and D. A. Salehah, "Demographic analysis of candidates supporter in twitter during indonesian presidential election 2019," in *2020 International Conference on ICT for Smart Society (ICISS)*, 2020, pp. 1–6.

[35] A. Suciati, A. Wibisono, and P. Mursanto, "Twitter buzzer detection for indonesian presidential election," in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2019, pp. 1–5.

[36] Q. You, S. Bhatia, T. Sun, and J. Luo, "The eyes of the beholder: Gender prediction using images posted in online social networks," in *2014 IEEE International Conference on Data Mining Workshop*, 2014, pp. 1026–1030.

[37] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2056–2067. [Online]. Available: https://doi.org/10.1145/3308558.3313684