Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 1, November 2023, pp. 79∼92
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i1.3318

Gender Classification of Twitter Users Using Convolutional Neural Network

Fitra Ahya Mubarok , Mohammad Reza Faisal , Dwi Kartini , Dodon Turianto Nugrahadi , Triando Hamonangan Saragih
Universitas Lambung Mangkurat, Banjarmasin, Indonesia

Article Info

Article history:

Received August 25, 2023
Revised September 14, 2023
Accepted October 03, 2023

Keywords:

Gender classification
Social media analysis
Twitter
Word2vec

ABSTRACT

Social media has become a place for social media analysts to obtain data to gain deeper insights and understanding of user behavior, trends, public opinion, and patterns associated with social media usage. Twitter is one of the most popular social media platforms where users can share messages or ”tweets” in a short text format. However, on Twitter, user information such as gender is not shown, but without realizing it or not, there is information about it in an unstructured manner. In social media analytics, gender is one of the important data that someone likes, so this research was conducted to determine the best accuracy for gender classification. The purpose of this study was to determine whether using combined data can improve the accuracy of gender classification using data from Twitter, tweets, and descriptions. The method used was word vector representation using word2vec and the application of a 2D Convolutional Neural Network (CNN) model. Word2vec was used to generate word vector representations that take into account the context and meaning of words in the text. The 2D CNN model extracted features from the word vector representation and performed gender classification. The research aimed to compare tweet data, descriptions, and a combination of tweets and descriptions to find the most accurate. The result of this study was that combined data between tweets and descriptions gets better accuracy 94,50% than using tweet data 91,60% or descriptions alone 88,10%.

Corresponding Author:

Mohammad Reza Faisal, +628122120152,
Faculty of Mathematics and Natural Sciences, Department of Computer Science,
Uiversitas Lambung Mangkurat, Banjarmasin, Indonesia,
Email: reza.faisal@ulm.ac.id.

How to Cite:
F. A. Mubarok, M. Reza Faisal, D. Kartini, D. T. Nugrahadi, and T. H. Saragih, ”Gender Classification of Twitter Users Using Convolutional Neural Network”, *MATRIK: Jurnal Managemen,Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 1, pp. 79-92, Nov, 2023.

1. INTRODUCTION

Social media is currently a trend in communication between people or marketing a product. Social media is an online platform where users can easily communicate, share, and create information through forums, wikis, and blogs. Almost everyone in the world uses social media to interact in cyberspace [1]. In the academic field, social media analytics for social media data has emerged, which includes collecting and analyzing various social media data and extracting valuable information [2]. Most data researchers need, such as personal information, preferences, and others, is easily obtained on social media. However, not all information can be obtained on certain social media such as Twitter, Instagram, TikTok, and others do not show gender information, so that when conducting social media analysis on these social media platforms, researchers do not know exactly how many men and women use social media because it is not given when accessing a Twitter account.

However, gender information is mostly provided consciously or unconsciously by users in an unstructured form [3]. The Twitter social media platform infers gender from various sources [4], such as usernames, screen names, descriptions, images, or user-generated content [3]. Based on previous research for gender classification using the BM25 feature extraction method and classification followed by the KNN algorithm method, the accuracy is 68.6%, with a total data of 1000 tweets on Twitter social media with 500 tweet accounts from male users and 500 from tweet accounts from female users [5]. For gender classification, the word2vec feature extraction method and classification using the CNN algorithm to get 91% accuracy with 22395 messages obtained from the Health Web Forum [6]. The research obtained an accuracy of 99% with TF-IDF feature extraction and CNN as an algorithm as its classification with a total data of 14166 Twitter tweets [7]. For gender classification using text data, the best result obtained is 57,14% with word2vec as feature extraction and logistic regression as a machine learning model [8].

For gender classification using N-gram feature extraction and Support Vector Machine classification algorithm with an accuracy of 93.2% with data totaling 65063 users where the data is in the form of profile picture, username, screen name, description, followers, and following, tweet, retweet, and favorite [3]. For comparing feature extraction between tf-idf and word2vec, using the XGBoost algorithm obtained an accuracy of 89.2% with tf-idf and 89.3% with word2vec [9]. Based on the above research, The difference between this research and previous research is gender classification by combining tweet data and descriptions in the Twitter user profile using word2vec and CNN for classification. The novelty of this research is using word2vec as word embedding and CNN as classification. This study aims to use methods from CNN and word embedding feature extraction, such as Word2vec, which is expected to provide greater accuracy than previous research.

2. RESEARCH METHOD

The stages of the research process in this study are described in a flow diagram, as shown in Figure 1. This research used a quantitative method, which used CNN as a classification algorithm and accuracy to determine the best result. The data source was from Twitter API on Rtweet to get data for this research.

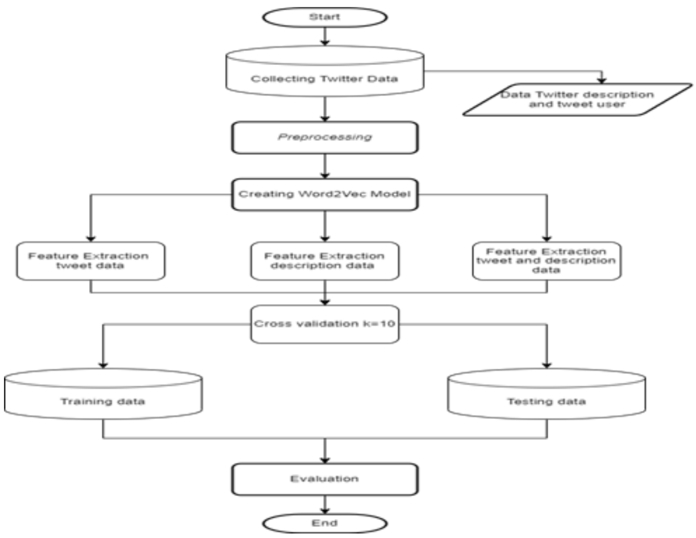

Figure 1. Research flow diagram

2.1. Collecting Twitter Data

The data in this study was taken using data scraping using the Twitter API on Rtweet with the search words ”sepak bola,” ”game,” and ”olahraga” and then retrieving user description data based on who did the tweet. Data scraping was the process of automatically extracting information from a website using a computer program or script. An example of the data collected is in Table 1.

Table 1. Twitter Data Collected

No.	Username	Description	Tweet
1	madeinheavenIV	*love is just a dumb luck*	*tidak ada pemain yang lebih besar dari sebuah club. tapi ini bukan tentang sepak bola hehe #Barcelona #Egypt*
2	l_yoshino2	⧫ IND/ENG ♀ she/her ◐ Hidup jalan terus tanpa henti	@bluelockfess Secara emang tahan untuk membuat pemain hebat/ maha karya pak ego, selain itu terikat keinginan kuat tentang sepak bola.
3	mstrprta	yesterday, today and tomorrow :)	MPV terus tapi losestreak 11 kali, kena troll terus, rusak ini game
4	radithisme	Sebuah seni tentang perjalan hidup, lewat tulisan. suka tidak suka hidup akan terus berjalan ke depan, meninggalkan kenangan menuju harapan baru.	@FWBESS Olahraga merakyat https://t.co/6BrZ0cA59l
5	sagigirllss	kadang receh kadang serius	Olahraga dulu kak biar kuat menjalani kehidupan https://t.co/reZNQ6pp4Z

2.2. Preprocessing

The collected data was not directly used in the data mining process using algorithms because it was still dirty data with much noise. So, preprocessing was needed to clean the data and select attributes that have little effect on the classification process. The preprocessing steps were as follows: Labeling is done on data extracted from Twitter [10]. The data was divided into two classes, namely male and female. Cleansing is the process of cleaning words that are not needed to reduce noise [11]. Case folding is a stage in preprocessing that aims to convert each word form into lowercase words [12]. Stemming is a preprocessing stage that aims to remove affixes contained in words or convert words into their basic form [13]. Stopword Removal or word filters are used to remove words contained in tweets and descriptions, but these words have no meaning or are not standardized (stopwords).

2.3. Creating Word2Vec Model

Data past the preprocessing stage was structured into a sentence matrix for classification. The process of converting words into matrices was done by making pre-trained word embedding in Indonesian with Word2vec. This document weighting process used Word2Vec. The vector dimension used was 100 dimensions. This model demonstrated the ability to learn patterns as linear relationships between word vectors. There are two word2vec algorithms, namely Continuous Bag-of-Word (CBOW) and Skip-gram [14]. In this research, the word2vec algorithm used was Skip-gram.

2.4. K-Fold Cross Validation

Before being entered into the model, the data was divided into training data and testing data using K-Fold Cross Validation with a value of K = 10 [15] so that the data was divided into 1800 training and 200 tests; with each iteration, there are 180 training and 20 testing.

2.5. Feature Extraction

Before being classified, the classification model of description data and tweets was transformed into a sentence matrix [10] for classification by weighting using the pre-trained word2vec that has been made. An example of feature extraction can be seen in Table 2.

Table 2. Example of feature extraction

sequence	V1	V2	...	V100
word1	numeric	numeric	numeric	numeric
word2	numeric	numeric	numeric	numeric
word3	numeric	numeric	numeric	numeric
word4	numeric	numeric	numeric	numeric
word5	numeric	numeric	numeric	numeric
word6	numeric	numeric	numeric	numeric

2.6. Creating Model Classification

The classification model created used the Convolutional Neural Network model for sentence classification, which refers to the architecture [16] with parameter adjustments. The parameters used in the CNN architecture [16] are word2vec dimension 300, Epoch 1-20, learning rate 0.001, dropout 0.5, and the optimization algorithm is adam. This parameter adjustment aimed to find the best parameter configuration so that the test could run optimally and produce classification performance with the highest value. The parameters adjusted in testing are the average input length [16], maximum input length [17], batchsize, and word2vec dimension size. The architecture of the CNN model can be seen in Figure 2. The classification process is as follows:

a. Input Layer
In the input layer, Twitter text data that has been reshaped into index form will be forwarded to the embedding layer to add vector values according to the predetermined dimensions of 100 dimensions. The size of the word length to be used, namely 12 and 49 are the maximum length for tweets, 8 and 30 are the maximum length for descriptions, and for combined tweet and description data, the length used is 20 and 79 for 1 channel while for 2 channels 12 and 49.

b. Embedding Layer
The embedding layer functioned to add vector values to the data as many as 100 dimensions so that there would be six types of input data for testing in the form of matrices measuring 12 x 100, 20 x 100, 49 x 100, 8 x 100, 30 x 100, and 79 x 100. The vector value was initiated according to the number of words in the data used for testing according to the word index in the pre-trained word embedding.

c. Convolutional Layer
In this convolutional layer, we used a filter dimension of 128, and the filter kernel used is 5x1 as in [16] and 5x5 due to using a kernel with the same dimensions as in [18]. The kernel (filter) is a small matrix that applies convolution operations to the input. After multiplication and summation operations between the filter weights and the weights of the input matrix and non-linear operations using the ReLU activation function, a feature map containing important features of lower dimensions in each hidden layer was obtained. Then, the feature map of each hidden layer became input to the max pooling layer in the next layer. This CNN model was formed using three convolutional layers.

d. Max Pooling Layer
The Max Pooling Layer took the highest value of the elements by means of a 2-dimensional window filter of size (5x 5) or (5x1), will traverse the matrix in a walk-through manner, and take the highest value of each 5x5 or 5x1 window traversed. Thus, each 5x5 or 5x1 window in the input data matrix will contribute one maximum value to the pooled output matrix, thus obtaining the most important information from the convolved feature map of each convolutional layer.

e. Fully Connected Layer
The output of the previously hidden layer, in the form of a feature map that had been reshaped into a vector, passed through the fully connected layer. The fully connected layer then performed mathematical operations between each neuron in this layer and the previous one. This allowed the neural network to learn more complex relationships between features in the input. After going through the fully connected layer, the output was connected to the output layer to classify the input based on the features that have been learned.

f. Output Layer
This layer used the softmax activation function and the Binary Crossentropy loss function. The softmax activation function and binary crossentropy loss function were used because two output variables represent multicategory classes, where the labels consist of 0 for the male class and 1 for the female class. The CNN classification model was trained using a batch size of 16 and 1-20 training epochs. The Adam optimizer function was used during training with a learning rate of 0.001. The purpose of using an optimizer was to reduce loss functions or errors that occur in neurons and filters in the model. In addition, the training applies the dropout regulation technique to overcome the overfitting problem that may occur during the training process [19].

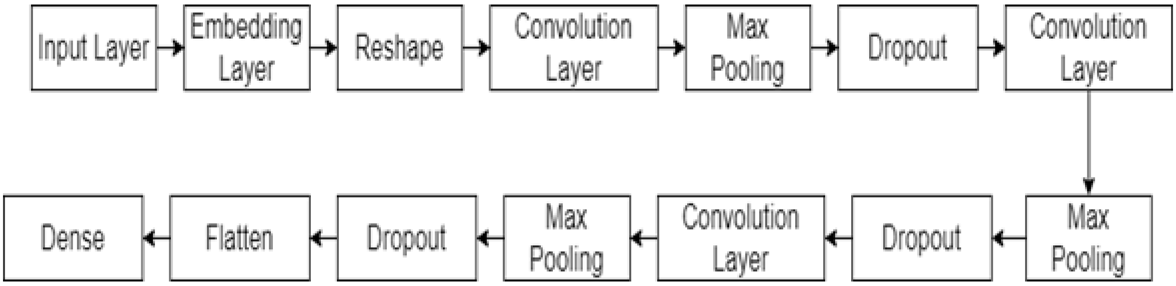

Figure 2. CNN Architecture

2.7. Evaluation

The evaluation was a stage to measure the performance results of the data mining techniques that have been carried out. To evaluate the classification model, we used the confusion matrix technique to calculate accuracy. Confusion matrix is a technique for measuring test results based on existing datasets. At the evaluation stage, accuracy was sought using the confusion matrix. The confusion matrix calculated metrics such as accuracy to evaluate the classification results of the method used. Accuracy is the proportion of true positive and true negative classification results [20].

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \times 100 \tag{1}$$

3. RESULT AND ANALYSIS

3.1. Collecting Twitter Data

Data obtained using the Twitter API shows that more than 5000 data were collected as a dataset. Of the 5,000 datasets, labeling was done until 2,000 data were obtained, which were classified into two categories: Male and Female, distributed evenly between 1,000 Males and 1,000 Females.

3.2. Preprocessing

a. Labeling
The data that has been collected was labeled manually according to their description and tweet or according to their Twitter account. For number dataset division based on manual labels shown in Table 3.

Table 3. Data Used

Data Class	Amount
Male	1000
Female	1000
Total	2000

1. Male
   Data is labeled as Male if:
   i. In the description, there was the word "he/him," "male," "boy," "dad," "father," or "man," as in the description of one user, "mom pick me up im scared ☹ | Male | alter" with the tweet "@milkymtcha_ Gerah soalnya pengen mandi tp olahraga dulu wkwkw " then this user is labeled Male because in the description there is the word Male.
   ii. In the description, written favorite game, favorite player, or favorite sports club as in the description of one user "Liverpool FC Outlet 23" with tweet @ligapurworejo @ligapurworejo Dulu SMP N 12 rata2 anak SSB Sawunggalih, Mitra Buana, & Bintang Laut, dilatih oleh Mas Widi (Sidarum) beliau staf TU kami yg bukan guru olahraga. Kalau SMP 4 kemungkinan anak2 IM Purworejo sama Bogowonto" then this user is labeled Male because, in the description, there was a favorite sports club.

iii. In checking the profile there is a photo of himself as in the description of one user " Co-Founder @fantasista_id|| Author of Pesta, Bola, dan Cerveja & Romantika Sepakbola " with the tweet " Terserah deh siapa pun Menpora-nya. Kuncinya, muda dan paham olahraga. Jabatannya menteri pemuda kok tua? Menteri olahraga kok cuman ngurusi sepak bola. Kalo itu sih Menbola. Kayak gak ada olahraga lain. Lebih perhatian deh ke olahraga prestasi, bola isinya kontroversi melulu.. https://t.co/5OmPUBfHdP" then in determining the label of this user must check the profile because cant be determined using only description or tweet, and this user has a photo of himself in his profile a man so that the label given is male.

2. Female

Data is labeled as female if :

i. In the description, there were words "she/her," "female," "girl," "mom," "mother," "daughter," or "woman," as in the description of one user, "Offgun's daughter" with a tweet @ubsansfess Pake koyo, trs istirahat dulu nder, tidur kek ato olahraga kecil" then this user is labeled a woman because in the description there is the word daughter.

ii. In the written description of the favorite idol, fandom, or shipper, as in the description of one user, "jgn tmenan sma aku klo km cuma reach out 1 ship like nomin only noren only please dni" with the tweet "ni cewe sok imut banget apaansih dah daritadi, ngerasa dispesialin bgt org mah olahraga mba biar tinggi https://t.co/MjLGC1GtFA" then this user is labeled a woman because in the description there is the word favorite idol.

iii. In the description, the word he/him was written. However, it was intended for other people, as in the description of one user "ILLUSIVE. Penned by Cthulhu. - Layung Jingga Artista. Not like the name, he is out of your expectation!" with the tweet "@ShoutOutOC Gue Layung kelahiran 97, mbak. Dibarengin olahraga aja biar makan banyaknya ttp jalan" then this user is labeled female because the description contains the word he but is shown for other people.

iv. In checking the profile, there is a photo of yourself as in the description of one of the users, "kadang receh kadang serius" with the tweet "Olahraga dulu kak biar kuat menjalani kehidupan https://t.co/reZNQ6pp4Z" then in determining the label of this user must check the profile, in this user there is a photo of herself, namely a woman so that the label given is a female.

b. Cleansing

An example of the results of applying cleansing to the data is presented in Table 4.

c. Case Folding

An example of the results of applying case folding to the data is presented in Table 4.

Table 4. Cleansing and Case Folding

Input Description	Output Description Cleansed and Case Folded	Input Tweet	Output Tweet Cleansed and Case Folded
Bahkan Pandawa pun lebih Kurawa ...#ForensicAuditor #SokGanteng #Sinetwit, #AManCalledAhok Movie	bahkan pandawa pun lebih kurawa forensicauditor sokganteng sinetwit amancalledahok movie	Zoey ; Tinggal menunggu 1 keterangan ahli utk nyatakan A layak dijadikan TSK.. Sudah itu kita fokus lagi ke Mario nya yah.. mau siapapun yang komporin.. si ngehek inilah yg paling biadab kelakuannya. Mudah2an teman selnya skr pencinta "salam olahraga" terbaik	zoey tinggal menunggu keterangan ahli utk nyatakan a layak dijadikan tsk sudah itu kita fokus lagi ke mario nya yah mau siapapun yang komporin si ngehek inilah yg paling biadab kelakuannya mudahan teman selnya skr pencinta salam olahraga terbaik
Offgun's daughter	offguns daughter	@ubsansfess Pake koyo, trs istirahat dulu nder, tidur kek ato olahraga kecil	ubsansfess pake koyo trs istirahat dulu nder tidur kek ato olahraga kecil

d. Stemming

Examples of the results of stemming application on the data are presented in Table 5.

e. Stopword Removal

Examples of the results of stopword application on the data are presented in Table 5.

Table 5. Stemming and Stopword Removal

Input Description	Output Description Stemmed and Stop-word Removal	Input Tweet	Output Tweet Stemmed and Stop-word Removal
bahkan pandawa pun lebih kurawa forensicau-ditor sokganteng sinetwit amancalledahok movie	pandawa,rawa,forensicauditor, sokgan-teng,sinetwit,amancalledahok,movie	Zoey ; Tinggal menunggu 1 keterangan ahli utk nyatakan A layak dijadikan TSK.. Sudah itu kita fokus lagi ke Mario nya yah.. mau siapapun yang komporin.. si ngehek inilah yg paling biadab kelakuannya. Mudah2an teman selnya skr pencinta ”salam olahraga” terbaik	zoey,tinggal,tunggu,terang,ahli,utk, nyata,a,layak,tsk,fokus,mario,nya, yah,komporin,si,ngehek,yg,biadab, laku,mudah,teman,sel,skr,cinta, salam,olahraga,
offguns daughter	offguns,daughter	@ubsansfess Pake koyo, trs istirahat dulu nder, tidur kek ato olahraga kecil	ubsansfess,pake,koyo,trs,istirahat,nder, tidur,kek,ato,olahraga

3.3. Creating Word2Vec Model

Word2vec feature extraction is useful for turning each word in the data into a vector. This research created a word2vec word embedding model with 100 feature dimensions. The Twitter data used amount of 2,000 description data, and the tweets used were also used for classification data with additional wiki data for the corpus, which had been preprocessed to clean the data. Then, a pre-trained Word2vec model with 100 dimensions was created using the gensim library in Python. The examples of word embedding models that have been generated are in Table 6.

Table 6. Pretrained Word2vec Model

word	index	V1	V2	...	V99	V100
satu	0	-0,46716	0,097488	...	0,282288	-0,6164
nol	1	-0,16638	0,223482	...	0,453218	-0,47889
dua	2	-0,03118	0,340899	...	0,624325	-0,59878
sembilan	3	-0,71234	0,558785	...	0,297537	-0,70694
dan	4	-0,48532	-0,50009	...	-0,14308	0,063257
yang	5	-0,42665	-0,53404	...	-0,28446	-0,14738
tiga	6	-0,74558	0,506237	...	0,346201	-0,54261
di	7	-0,26216	-115,897	...	0,881984	0,661438
empat	8	-0,60675	0,593217	...	0,424913	-0,55264
lima	9	-0,62302	0,503963	...	0,461822	-0,41288
...	...	...	...	...	...	...
xaveriana	703745	0,066068	-0,00856	...	0,004858737	0,086722992

3.4. Feature Extraction

After the pre-trained word embedding model using Word2Vec was successfully created, the next process was feature extraction using word2vec. The function of this feature extraction was to make the data into a vector form according to the weighting in the word2vec that had been made. This weighting occurred in the embedding layer before insertion into the convolution layer. The features obtained matched the number of dimensions in word2vec, which was 100 dimensions. An example of extracting features can be seen in Table 7 with the tweet data of one of the users, ”sheichanakemi serius gak tau game horor multiplayer” with a maximum input of 12.

Table 7. Example of Feature Extraction

sequence	word	V1	V2	...	V100
1		0	0	...	0
	...	...	...	...	...
6	sheichanakemi	0	0	...	0
7	serius	0,049925	-0,04022	...	-0,26696
8	gak	0,237407	-0,10593	...	-0,14467
9	tau	0,228046	-0,15588	...	0,068024
10	game	-0,15548	-0,2642	...	0,317519
11	horor	0,024386	-0,21679	...	0,234644
12	multiplayer	-0,26849	-0,64631	...	0,73904

3.5. Creating Classification Model

After feature extraction using Word2Vec is successfully made, the next process in the research is to create a Convolutional Neural Network classification model. The architecture used for the classification model refers to [16] with modifications in some of its parameters to get the best accuracy value. The classification model is made using the Python package, namely Keras Tensorflow. The model uses word2vec with a dimension of 100. The architecture that has been created can be seen in Table 8.

Table 8. Pembagian data untuk Training dan Testing

Layer	Output Shape
Input Layer	[(None, 49, 1)]
Embedding	(None, 49, 1, 100)
Reshape	(None, 49, 100, 1)
Convolution Layer	(None, 25, 50, 128)
Max Pooling	(None, 13, 25, 128)
Dropout	(None, 13, 25, 128)
Convolution Layer	(None, 7, 13, 128)
Max Pooling	(None, 4, 7, 128)
Dropout	(None, 4, 7, 128)
Convolution Layer	(None, 2, 4, 128)
Max Pooling	(None, 1, 2, 128)
Dropout	(None, 1, 2, 128)
Flatten	(None, 256)
Dense	(None, 2)

3.6. Classification And Evaluation

a Tweet

The results of using tweet data with cnn can be seen in Table 9, which is divided into 4, namely the difference in kernel size and input size, with the difference for kernel size 5x1 and 5x5 and input size 12 and 49. The largest result obtained on tweet data was 91.60% with input size 49 and kernel size 5×5, as shown in Figure 3.

Table 9. Result for Tweet Data

Epoch	Kernel Size	Input Size	Accuracy
18	5x1	12	89,45%
18	5x5	12	91,10%
20	5x1	49	89,35%
20	5x5	49	91,60%

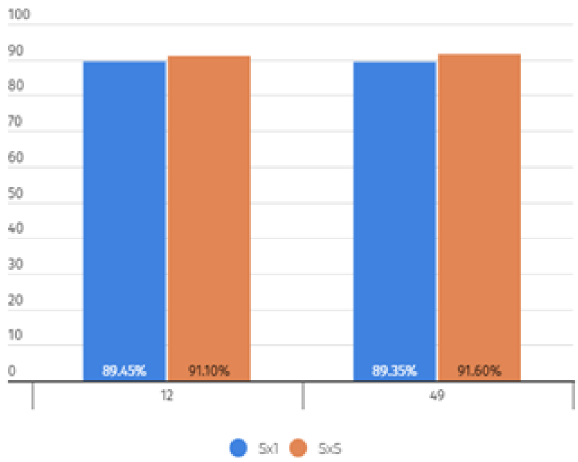

Figure 3. Comparison Tweet Accuracy

b Description
The input sizes, which were 8 and 30, are different on description data with the same difference as tweet data. The results can be seen in Table 10. The largest result obtained on description data is 88.10% with input size 30 and kernel size 5 × 5, as shown in Figure 4.

Table 10. Result for Description Data

Epoch	Kernel Size	Input Size	Accuracy
20	5 ×1	8	84,50%
19	5 × 5	8	87,70%
19	5 × 1	30	86,65%
20	5 × 5	30	88,10%

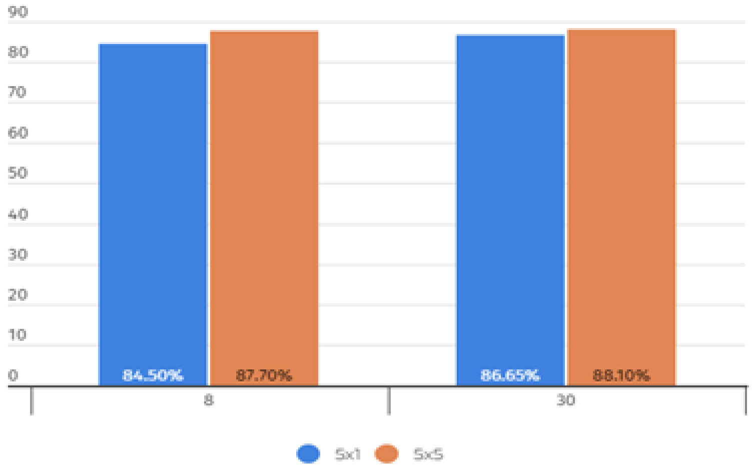

Figure 4. Comparison Description Accuracy

c Tweet And Description in the combined tweet and description data, there are differences in combining the data using only 1 channel and 2 channels. For channel 1, the input sizes used were 20 and 79, while for channel 2, the input sizes used were 12 and 49, as shown in Table 11. The highest result for the combined tweet and description data was 94.50% using 1 channel and input sizes 20 and 79, as shown in Figure 5.

Table 11. Result for Data Combined

Epoch	Kernel Size	Input Size	Data	Accuracy
20	5 × 1	20	Description dan Tweet 1 channel	94,50%
19	5 × 5	20	Description dan Tweet 1 channel	93.15%
19	5 × 1	79	Description dan Tweet 1 channel	94,50%
20	5 × 5	79	Description dan Tweet 1 channel	93,05%
20	5 × 1	12	Description dan Tweet 2 channel	91,95%
19	5 × 5	12	Description dan Tweet 2 channel	92,90%
20	5 × 1	49	Description dan Tweet 2 channel	93,50%
20	5 × 5	49	Description dan Tweet 2 channel	93%

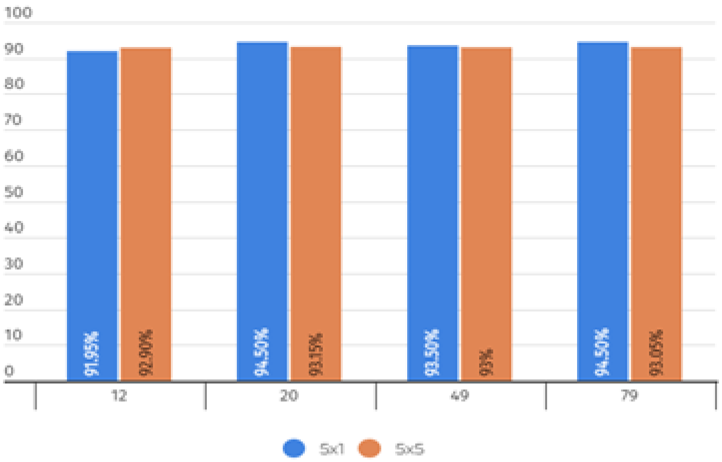

Figure 5. Comparison Combined Tweet and Description Accuracy

3.7. Analysis

Each of the best results on all data can be seen in Table 12. The highest accuracy result was the combined data between descriptions and tweets, with a result of 94.50%, as shown in Figure 6. For accuracy, using epochs from 1st to 20th can be seen in Figure 7, Figure 8, Figure 9, and Figure 10, which different sizes of kernel used for classification.

Table 12. Result for All Data

Epoch	Kernel Size	Input Size	Data	Accuracy
20	5×5	49	Tweet	91,60%
20	5×5	30	Description	88,10%
20	5×1	20	Description dan Tweet	94,50%
19	5×1	79	Description dan Tweet	94,50%

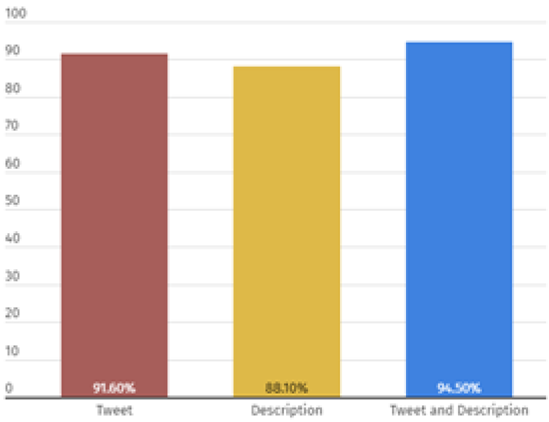

Figure 6. Comparison Result for All Data

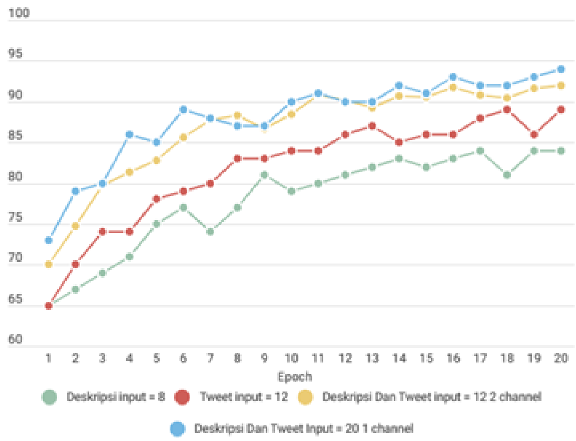

Figure 7. Accuracy Each Epoch Using Kernel Size 5 × 1

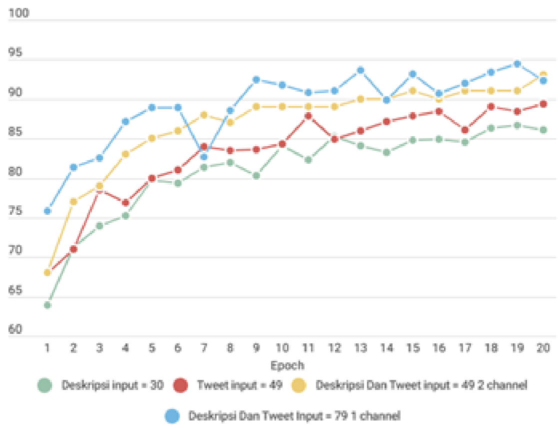

Figure 8. Accuracy Each Epoch Using Kernel Size 5 × 1

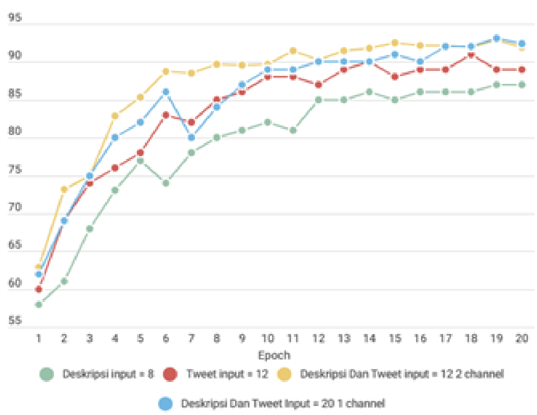

Figure 9. Accuracy Each Epoch Using Kernel Size 5 × 5

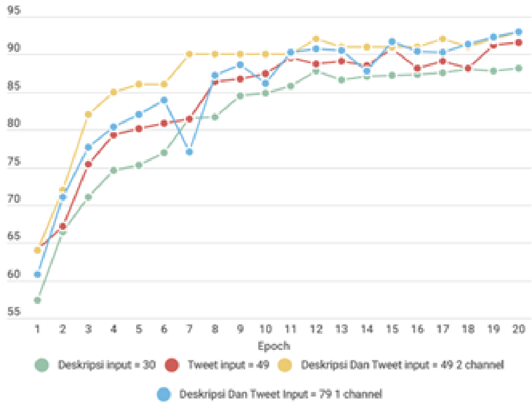

Figure 10. Accuracy Each Epoch using Kernel Size 5×5

Based on the accuracy value, it can be concluded that for gender classification using data on Twitter, users with the highest accuracy were data with a combination of descriptions and tweets using 1 channel on CNN. This may be because descriptions and tweets have more information and features when entered into the model compared to using only description or tweet data. In contrast, in the use of channels, the highest accuracy was 1 channel because, by using 1 channel, the data was combined, and the number of words was adjusted to each input so that there was not too much padding and more information. The model could combine the two inputs to extract better and more accurate patterns to improve classification performance and accuracy. The best result in this research obtained was 94,50%, obtained better result than research from [8] 57,14%, and research from [3] 93,2%, which aligns with this research using combined data to classify gender.

3.8. Discussion

Based on the test results, the kernel size was very influential in getting the highest accuracy, with the highest accuracy being the combined data between descriptions and tweets at 94.50%. The description and tweet data alone were more accurate using a larger kernel size. However, in the combined data, the accuracy was better using a 5x1 kernel size. This may happen because each data has more unique characteristics when combined, so that with a smaller kernel size, it gets clearer feature information, resulting in an increase in accuracy. It also happens because the 5x1 kernel size uses 1x1 strides while 5x5 uses 2x2 strides so that the 5x1 kernel gets more information for the combined data because there is no dimensional reduction when using stride 1x1, using 1 channel to do the merging input process at the beginning while for 2 channels to do the merging process before the output so that things like this also affect getting information on the data presented.

In Figure 7, Figure 8, Figure 9, and Figure 10, some graphs go up and down. This is because when randomized data classification was carried out using k-fold cross-validation at the beginning, the algorithm was confused in determining men and women so that fold 1, 2, or 3 get lower accuracy than the previous epoch so that there was a decrease in accuracy but to see if there was true overfitting because previously there were graphs up and down and the highest result is not epoch 20 then the research was carried out adding 5 epochs so that it reached epoch 25 can see in figure 11. In Figure 11, the results obtained did not increase the highest result in the previous one, so it can be said that there was a possibility of overfitting.

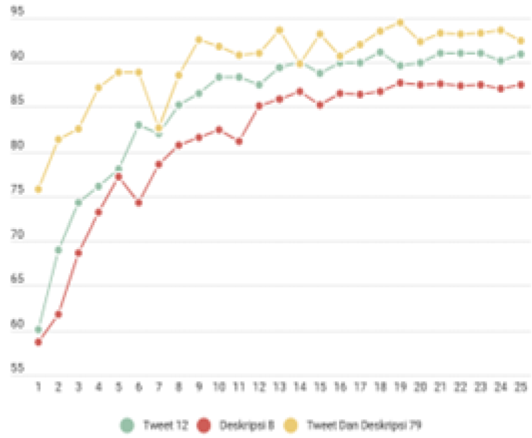

Figure 11. Accuracy Epoch Until Epoch 25

4. CONCLUSION

The results of gender classification accuracy using tweet data only got a result of 91.60%, and the input length used was the length of the tweet data, which was 49, with the kernel size used being 5x5. The results of gender classification accuracy using description data only got 88.10% lower than using tweet data only with an input length using the maximum length of description data, which was 30 and kernel size 5x5. Finally, the results of gender classification accuracy using combined data between tweets and descriptions of 94.50% showed a significant increase in accuracy results compared to using tweet data alone or only with description data alone with input length using input sizes 20 and 79 with differences in convergence on epochs with input size 20 reaching convergence at epoch 20 and 79 at epoch 19 with kernel size 5x1. The result obtained from this research was using combined data from Twitter to classify gender better than only using one data such as tweet and description, using word2vec and cnn also obtained better results than previous research. In future research, the researcher can use different word embedding methods such as gloves or fasttext to obtain better accuracy.

5. ACKNOWLEDGEMENTS

The authors would like to thank everyone who helped with this work, especially the anonymous reviewers, the chief/managing editors, and the Matrik: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer staff.

6. DECLARATIONS

AUTHOR CONTIBUTION

This research was conducted by five authors with a division of tasks: Fitra Ahya Mubarok took care of data collection and analysis; Mohammad Reza Faisal worked on ideas, research flow, and design; Dwi Kartini wrote articles and checked the data results; Dodon Turianto Nugrahadi provided advice and criticism about the research; Triando Hamonangan Saragih tested the research models.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

I have no declaration under financial, general, and institutional competing interests.

REFERENCES

[1] U. Sivarajah, Z. Irani, S. Gupta, and K. Mahroof, “Role of big data and social media analytics for business to business sustainability: A participatory web context,” *Industrial Marketing Management*, vol. 86, no. April, pp. 163–179, apr 2020.

[2] J. Choi, J. Yoon, J. Chung, B. Y. Coh, and J. M. Lee, “Social media analytics and business intelligence research: A systematic review,” *Information Processing and Management*, vol. 57, no. 6, pp. 1–18, nov 2020.

[3] M. Vicente, F. Batista, and J. P. Carvalho, "Gender detection of Twitter users based on multiple information sources," in *Studies in Computational Intelligence*. Springer Verlag, 2019, vol. 794, pp. 39–54.

[4] E. Fosch-Villaronga, A. Poulsen, R. Søraa, and B. Custers, "A little bird told me your gender: Gender inferences in social media," *Information Processing & Management*, vol. 58, no. 3, pp. 1–13, may 2021.

[5] A. Selma Zakia, "Klasifikasi Jenis Kelamin Pengguna Twitter dengan menggunakan Metode BM25 dan K-Nearest Neighbor (KNN)," Tech. Rep. 10, 2020.

[6] S. Park and J. Woo, "Gender classification using sentiment analysis and deep learning in a health web forum," *Applied Sciences (Switzerland)*, vol. 9, no. 6, pp. 1–12, 2019.

[7] R. Alroobaea, S. Alafif, S. Alhomidi, A. Aldahass, R. Hamed, R. Mulla, and B. Alotaibi, "A Decision Support System for Detecting Age and Gender from Twitter Feeds based on a Comparative Experiments," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 370–376, dec 2020.

[8] P. Vashisth and K. Meehan, "Gender Classification using Twitter Text Data," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, jun 2020, pp. 1–6.

[9] I. R. Hendrawan, E. Utami, and A. D. Hartanto, "Analisis Perbandingan Metode Tf-Idf dan Word2vec pada Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal di Indonesia," *Smart Comp*, vol. 11, no. 3, pp. 497–503, 2022.

[10] M. R. Faisal, M. I. Mazdadi, R. A. Nugroho, F. Abadi, and Others, "Eye Witness Message Identification on Forest Fires Disaster Using Convolutional Neural Network," *Journal of Data Science and Software Engineering*, vol. 2, no. 2, pp. 100–108, 2021.

[11] K. Y. Firlia, M. R. Faisal, D. Kartini, R. A. Nugroho, and F. Abadi, "Analysis of New Features on the Performance of the Support Vector Machine Algorithm in Classification of Natural Disaster Messages," in *Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, 2021, pp. 317–322.

[12] M. Rusli, "Ekstraksi Fitur Menggunakan Model Word2vec pada Sentiment Analysis Kolom Komentar Kuisioner Evaluasi Dosen oleh Mahasiswa," *KLIK - Kumpulan Jurnal Ilmu Komputer*, vol. 7, no. 1, pp. 35–47, mar 2020.

[13] M. Padhilah, D. Kartini, and D. T. Nugrahadi, "Implementasi Neural Network Multilayer Perceptron Dan Stemming Nazief & Adriani Pada Chatbot Faq Prakerja," *Jurnal Sains Komputer & Informatika (J-SAKTI)*, vol. 6, no. 2, pp. 671–685, 2022.

[14] A. Nurdin, B. Anggo, S. Aji, A. Bustamin, and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2vec, Glove, dan Fasttext pada Klasifikasi Teks," *Jurnal Tekno Kompak*, vol. 14, no. 2, pp. 74–79, 2020.

[15] L. Islami, I. Budiman, M. R. Faisal, and F. Abadi, "Prototype Generation Berdasarkan Geometric Mean Untuk Data Reduction pada Algoritma K Nearest Neighbour," *Jurnal Data Science & Informatika (JDSI)*, vol. 2, no. 2, pp. 53–59, 2022.

[16] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "the Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (CNN) Text Classification," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 2, pp. 349–359, 2022.

[17] J. Bai, I. Shim, and S. Park, "MEXN: Multi-Stage Extraction Network for Patent Document Classification," *Applied Sciences*, vol. 10, no. 18, pp. 1–14, sep 2020.

[18] N. Ketkar and J. Moolayil, *Deep Learning with Python*. Berkeley: Apress, 2021.

[19] G. S. Nandini, A. S. Kumar, and C. K, "Dropout technique for image classification based on extreme learning machine," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 111–116, 2021.

[20] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images," 2020.