




Machine learning based approaches for age and gender prediction from tweets

Rishabh Katna¹ · Kashish Kalsi¹ · Srajika Gupta¹ · Divakar Yadav¹  · Arun Kumar Yadav¹

Received: 18 May 2021 / Revised: 17 January 2022 / Accepted: 9 March 2022 /
Published online: 29 March 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Author profiling is the process of analysing textual data to extract information about various personality traits of the author. It has both commercial and social implications. Many approaches have been proposed in past to increase the accuracy of the extracted information. In this paper, we applied natural language processing and machine-learning approach for author profiling. NLP techniques like Tokenization, lemmatization, word and char n-grams are used in integration with machine learning classifiers like logistic regression (LR), random forest (RF), decision tree (DT) and support vector machine (SVM). The proposed method obtained an accuracy of 81.2%, 79.8%, 63.2% and 88.0% for the four classifiers respectively for gender prediction and 72.5%, 68.1%, 53.7% and 81.0% respectively for age prediction i.e. SVM outperformed the other classifiers with an accuracy of 88.0% for gender prediction and 81.0% for age prediction.

Keywords Twitter dataset · Author profiling · NLP · Age prediction · Gender prediction · TF-IDF

✉ Divakar Yadav
divakaryadav@nith.ac.in

Rishabh Katna
rishabhkatna2228@gmail.com

Kashish Kalsi
kalsi.kashish@gmail.com

Srajika Gupta
srajika49@gmail.com

Arun Kumar Yadav
ayadav@nith.ac.in

¹ Department of Computer Science & Engineering, NIT Hamirpur, Hamirpur (HP) 177005, India

1 Introduction

The engagement of users on social media is increasing day by day. According to a survey by Dave Chaffey in 2017 [32], 53.6% of the world's population uses social media and the average daily social media usage of internet users worldwide amounted to 2 h 25 min per day. Today social media websites are getting millions of active users daily. Social media like Instagram, Facebook, etc. serve 500 million active users daily. This rise in the popularity of social media usage has led to the age of data. These social media websites collect lots of user data and are using the data to improve the user experience. The focus of these websites has shifted to the personalization of user experience, thus the author profiling based on textual data has been a topic of interest for researchers.

Author profiling is analysing textual data to extract information about various personality traits and characteristics of authors. The most common features are the age and gender of the author. In addition, personality traits like a sense of humour, fields of interest can be very beneficial for some studies. In fact, author profiling is one of the fields in automatic authorship identification (AAI). The first use of author profiling dates back to the nineteenth century when Mendenhall [14] applied this process to the most popular authors, Francis Bacon, William Shakespeare and Christopher Marlowe. This helped to know about their stylistic differences by analysing the average word lengths of the authors.

Today author profiling has commercial uses as well. Companies are using the data collected by the users to understand their user base in a better way. Companies like Google and Facebook generate personalized ads using this information. Other companies use this information to give personalized offers such as loan offers, clothes, gadgets and so on. This information generates billions of dollars for these companies every year. Companies can use this information to detect fraud and criminal activities. Hence, author profiling has high commercial values.

Adding to the commercial applications, it has social applications as well. A rise in social media users also gave rise to online criminal activities. Criminals posing as teenagers can conduct criminal activities like selling illegal products, paedophilia etc. through these platforms and will go undetected because there is no way to verify whether the information given by the user is accurate or not. Author profiling can solve this problem. By analysing the posts and chats of users, information about age and gender can be predicted and using this information, suspects can be identified, which may help stop illegal activities and crimes [3].

The study of author profiling on computer-based communication methods started in the '90s and early 2000s. Papers like [1–4] did make significant contributions in this field. Herring & Paolillo [2] proposed a detailed analysis of the topics and language used by men and women in different communication modes as follows:

- **Spoken:** In the spoken mode of communication, females tend to talk more about relationships and internal states [18] whereas males tend to talk more about objects such as cars, computers, sports, politics etc. [19].
- **Written:** Language used in written mode such as books, articles etc. is more formal than spoken language. The style followed by females tends to be more gossipy and they generally follow a diary writing style [20]. However, scientific writing was considered a more masculine trait [21].
- **Computer-Based:** The computer-based communication is more written (chats) and the style followed is more informal. Here women use a more polite, emotionally expressive

and less verbose language whereas men chose a more informational and factual language [22].

Although these studies were conducted in the late 90's and early 2000s, they still gave us valuable information about the nature of discussions between men and women. Although the platforms have changed, personalities and traits are still very similar. Hence, these observations proved to be very beneficial to modern research on this topic. One of the earliest studies in this field was by Mendenhall [14] and made a major contribution to this field. Modern researchers conducted their studies using different datasets. [2, 3] predicted gender on online social networks. In that approach, selected blogging websites for their studies. Authors of [4, 5, 7, 11] worked on the Twitter dataset.

In this work, the basic objective is to improve the accuracy to classify gender and age from the tweet dataset. After a detailed study of existing state-of-the-art methods, we found that Daneshvar et al. [7] got the highest accuracy among all the others. So, we tried different classifiers along with improvised data pre-processing steps by adding regex-based data cleaning and lemmatization using nltk's WordnetLemmatizer. We also changed the dataset proportions between training and testing. The above changes ended up producing higher accuracy than the other studies.

The rest of the article is organized as follows: In section 2, we describe previous studies performed in this field. In section 3 we discuss in brief about the dataset. In section 4, we present our approach for age and gender prediction, detailed discussion about pre-processing, feature extraction and classification. In section 5, we discuss the results obtained by the implementation of our approach on PAN at CLEF 2015 [24] author profiling task dataset. Section 6, contains the conclusion and future scope of the study. Table 1, given below lists some of the abbreviations used in the article.

2 Related work

Author profiling can be very beneficial and this is the reason researchers all around the world are contributing to this field. Many researchers have made significant contributions to the domain of profiling authors [1–5, 7, 10, 11, 14]. One of the earliest studies in this field was conducted by Mendenhall [14]. He wanted to identify the author of a book, play or poem by analysing the mean word length used by the author. He built a data set based on works by popular authors such as Francis Bacon, William Shakespeare and Chris Marlowe. He formed a

Table 1 Abbreviation and Full forms

Abbreviation	Full Form
TF-IDF	Term frequency-Inverse document frequency
RF	Random forest
SVM	Support vector machine
LR	Logistic regression
DT	Decision tree
API	Application programming interface
CNN	Convolutional neural network
POS	Part of speech
LIWC	Linguistic inquiry and word count

‘word spectrum’ for each author by keeping word lengths at the x-axis and word frequency at the y-axis. Looking at the curve, he could find the author of a poem by examining the “spectrum of words” in the book. Though today we have available better methods for feature extraction like n-grams still, his work set a path for the coming generations of scientists.

Work done by Herring et al. [17] was based on the effect of gender on the genre of the blogs. Herring & Paolillo [2] studied the effects of gender on the writing styles of the authors. They collected 100 random blogs on the Web for their research and extracted style-based features based on different writing habits of men and women. Further, they used logistic regression to classify blogs but found no meaningful relationship between the gender and gender of blogs. The limitation of this study was that they used a small dataset. They also conducted the study on a formal language dataset; therefore, these techniques may not be very effective on informal data.

Argamon et al. [1] took an unbiased approach. They changed the problem statement to “given a set of written documents with the author’s gender labelled, predict the gender of an unknown document”. They used a dataset from British National Corpus. The dataset had 604 documents. They used the EG algorithm [13] as a Classifier. EG algorithm is a generalization of the Balanced Winnow [12]. The algorithm extracted 50 features from the dataset. The dominant masculine characteristics included determinants (a, an, etc.) and quantifiers (one, two, plus, etc.).

The dominant feminine features included personal pronouns (I, you, her, herself etc.). They could predict the gender with an accuracy of 80%. Nguyen et al. [15] took a multi-corpora approach. They extracted three corpora: blog corpus, transcribed telephone speech [23], and posts from online forums on breast cancer. Further, they extracted the following features from the dataset:

- Unigrams.
- POS unigrams and bigrams.
- LIWC [24].

They took a different approach to classification. Rather than taking age as a discrete variable as in [1–5], they took age as a continuous variable. They used a linear regression classifier. They have a correlation of 0.74 and MAE of 4.1 to 6.8 years. Some studies suggested that style-based features are better for predicting age and content-based features are better for gender prediction as males and females have different writing patterns and there are differences in topics of interest [9]. These studies proposed some effective approaches towards this problem, but they were more focused on formal text and communication. Our study is more inclined towards the informal text.

Peersman et al. [3] collected their dataset from a Belgium based social media platform called Netlog. They extracted 1.5 million posts containing 18 million tokens. They applied the chi-square feature extraction metric. From the extracted features, unigrams, bigrams, trigrams and tetragrams were used for classification. They used SVM classifiers and got an accuracy of 81.3% for age. For age and gender combined, they got an accuracy of 66.3%. The low accuracy values are due to the small size of the dataset. However, they worked on non-formal language and hence their techniques were very helpful in our study.

Burger et al. [4] worked on the Twitter corpus. The main focus of their study was to test the changes of a very large corpus on the classification results. They extracted 213 million tweets from 18.5 million users using twitter’s API. They also extracted the author’s full name, screen name, self-description in addition to the tweets. Similar to Nguyen et al. [15], they also used the word and char n-grams as characteristics. They used three different classifiers: SVM,

Naive-Bayes and Balanced Winnow [12]. SVM gave an accuracy of 71%, Naive-Bayes 67% and Balanced Winnow 74%.

Divakar Yadav et al. [30] proposed a method for prediction of age groups namely teenagers, adults and senior citizens from textual data collected from twitter using different classifiers such as K-Nearest Neighbor (KNN), Multi-layer Perceptron (MLP), Decision tree, Random Forest and Support Vector Machine (SVM). Among all, random forest performance was found better than other classification methods for each age group.

Daneshvar et al. [7] used the PAN2018 dataset [6]. The dataset consisted of 300,000 tweets from 3000 users. Just like Burger et al. [4], they decided to extract language-independent features from the dataset, as they tend to produce better results than stylistic features [5]. They used a Tf-Idf vector of word and char n-grams combined as their feature set. They used three different classifiers: Naive-Bayes, SVM and logistic regression. SVM produced an accuracy of 81.8%. The main limitation of their study was that they did not use very effective pre-processing and data cleaning methods. Overall, their approach produced better results than previous studies, hence our study builds upon their work and tries to overcome the limitations of this study by using better methodologies.

Sezerer et al. [10] used various classifiers such as KNN, Naive-Nayes, covering rules, and backpropagation to predict gender and age on chat messages. Most age and gender predictions rely on n-gram features. Neural network-based models were also proposed. They were able to extract both syntactic and semantic features. An improved version of RNN + n-grams also produced better results.

Aroju et al. [11] used the n-gram feature, also one feature Linguistic Inquiry and Word Count (LIWC) [24] for personality prediction and semantic analysis. After extracting the word n-gram features, a linear model with stochastic gradient descent (SGD) learning was used. A concise summary of works done in the domain is given in Table 2.

3 Dataset

The PAN [25] author profiling dataset is being used for this study. In original dataset, 230 authors' tweets are included in the training dataset. In addition, 57 users' tweets are included in the test dataset as shown in Table 3.

Total of 100 tweets and 10 images are provided for each author. In this work, only the tweets of the authors have been used and not the images as our work is related to text processing. Also, some tweaks are made in the dataset to get better accuracy i.e. combining the dataset into 1 and then split into 2 parts in a ratio of 1:9 (test dataset: training dataset) in order to increase the size of the training dataset to 258 and test dataset to 29 as shown in Table 4.

Each author's tweets are stored in an XML file. The XML files were named after the id of the author as "author_id.xml". The format of data in each XML file is shown in Figure 1. Each document in the document tree is a tweet posted by the user.

4 Proposed architecture of the model

The proposed model of age and gender prediction passes through several steps as shown in Fig. 2. The initial step it to pre-process the collected raw data followed by extraction of useful

Table 2 Brief summary of the state-of-the-art methods

Article	Year of Publication	Dataset	Methodology	Performance measure
Argamon et al. [1]	2003	British National Corpus	EG Algorithm [12]	Accuracy: 80%
Herring et al. [2]	2006	100 weblogs	Logistic Regression	–
Peersman et al. [3]	2011	Netlog corpus	S.V.M	Accuracy: Age: 71.3%; Age and gender: 66.3%
Burger et al. [4]	2011	Twitter Dataset	S.V.M, Naive-Bayes, Balanced-Winnow [13]	Accuracy: SVM: 71% Naive-Bayes: 67% Balanced-Winnow: 74% Accuracy: 81.8%
Daneshvar et al. [7]	2018	PAN2018	Naive-Bayes, SVM, Logistic-Regression	Accuracy: NB-Dtree: 78.3%(with 5-g)
Alowbidi et al. [8]	2013	Twitter Dataset	Naive-Bayes, Decision-Tree, NB-DT Hybrid	Was able to distinguish authors based on resulting word spectrum
Mendenthal T. [14]	2011	Works of popular authors like Shakespeare, Marlowe and Bacon	Word frequency and word spectrum	Correlation: 0.74; MAE: 4.1 to 6.8 years
Nguyen et al. [15]	2011	Three corpora: 1. Blog corpus 2. Transcribed phone speech. 3. Online Posts	Linear Regression	
Ljuberic et al. [5]	2017	Twisty Corpus [16]	SVM on 5 g	F1 Score: 69.92%
Sezerer et al. [9]	2018	PAN2018	CNN	Accuracy: 75.1%
Sezerer et al. [10]	2019	PAN2018	CNN and RNN	Accuracy: CNN: 78.47; RNN: 82.31
Rexha et al. [26]	2018	Reuters Corpus Volume 1 (RCV1)	SVM with Information Gain	Accuracy: 73%

Table 2 (continued)

Article	Year of Publication	Dataset	Methodology	Performance measure
Okuno et al. [29]	2014	Twitter Dataset(10,000 tweets)	char-ngrams	Precision@1: 0.532
Rangel et al. [29]	2013	SMS Dataset	Euclidean Distance, Cosine Distance using unigrams	–
Alowbidi et al. [28]	2013	Twitter Dataset	PNN, Decision Tree, Naive-Bayes, NB-DTree Hybrid	Accuracy: PNN: 71.6%; NB-DTree: 70.7%
Rao et al. [27]	2010	Twitter Dataset	SVM on N-grams and Stacked model	Accuracy(max): Age: 74.11%(stacked model) Gender: 72.33%(stacked model)

Table 3 PAN 2015 original dataset

Language	Train dataset size	Test dataset size
English	230×100 (Avg. Tweets)	57×100 (Avg. Tweets)

features, and then train the dataset using various classifiers, further applying different techniques to make the model more accurate and effective for unseen data. Thus, the model consists of three phases:

- Pre-processing
- Feature Extraction
- Classification

Detailed explanation of the model is given in the following sections.

4.1 Pre-processing

The most important step in text analysis is pre-processing the input text. The steps for pre-processing tweets are described below and the flow diagram is shown in Fig. 3:

- The initial step to process the data is to lowercase all the words as it helps to get rid of unhelpful parts or noise of the data and maintain consistency.
- As the model is for the Twitter dataset, in the next step, we need to remove embedded # mentions, @ mentions and URL mentions as they are not very valuable for our study and can negatively affect the results.
- In the next step, remove slang words and replace the characters with a frequency greater than three with fewer characters to intact the same meanings of the words. e.g. the word “Brooooo” will become “Broo”.
- The next step is to tokenize all tweets into words and remove punctuation as they are irrelevant and can negatively bias results. Split the tweet into different tokens and store them in a list. This list represents the Bag of Words.
- The next step is lemmatization. Applying lemmatizer to each token in the bag of tokens, the pos tag chosen is ‘a’(Adjective). Lemmatization helps to achieve the root forms (synonyms) of inflected words.
- Removal of the stopwords from lemmatized tokens. Stop words are the words that occur very frequently in a sentence but do not express the sentiment of the author. E.g. ‘it’, ‘is’, ‘a’, ‘us’, ‘this’ etc. Stopwords are used very frequently hence it can negatively affect the performance of a model. Therefore, it is necessary to remove them.
- The bag of tokens are joined to form a sentence and thus the pre-processing is completed.

Table 4 PAN 2015 tweaked dataset

Language	Train dataset size	Test dataset size
English	258×100 (Avg. Tweets)	29×100 (Avg. Tweets)

Fig. 1 Format of each XML file

```

    <author lang="en">
    <documents>
    ...
    </documents>
    </author>
    
```

The lowercasing, removal of #mention, @mention and URL are done using a function called ‘clean’ which is presented in the module ‘preprocessor’. Tokenization of the tweets is implemented using nltk’s TweetTokenizer() which is found in the module ‘nltk.tokenize.casual’. Lemmatization is performed using nltk’s WordNetLemmatizer() which is found in the module ‘nltk.stem’. Gensim’s remove_stopwords function is used to remove stopwords from the tweets, which is found in the module ‘gensim.parsing.preprocessing’ and thus the pre-processing is completed.

4.2 Feature extraction

The task of author profiling of tweets has numerous applications and is a well-researched topic. Feature extraction is the most essential phase of a classification model.

N-gram is a contiguous sequence of n words collected from the tweet. Character n-grams and word n-grams were used as a feature set in [4, 8]. Similar to Daneshvar et al. [7], we have used Tf-Idf for term scoring. We have used Word n-grams of (1, 3) and character n-grams of (2, 5) and

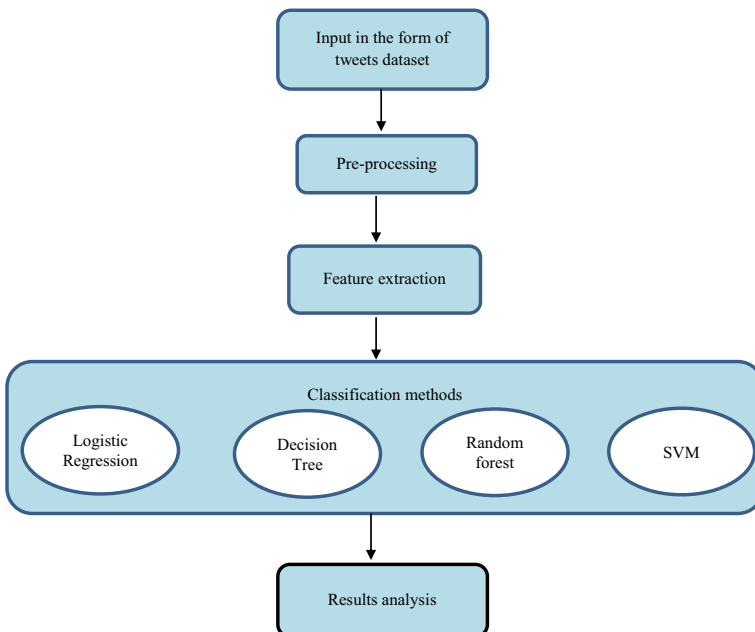
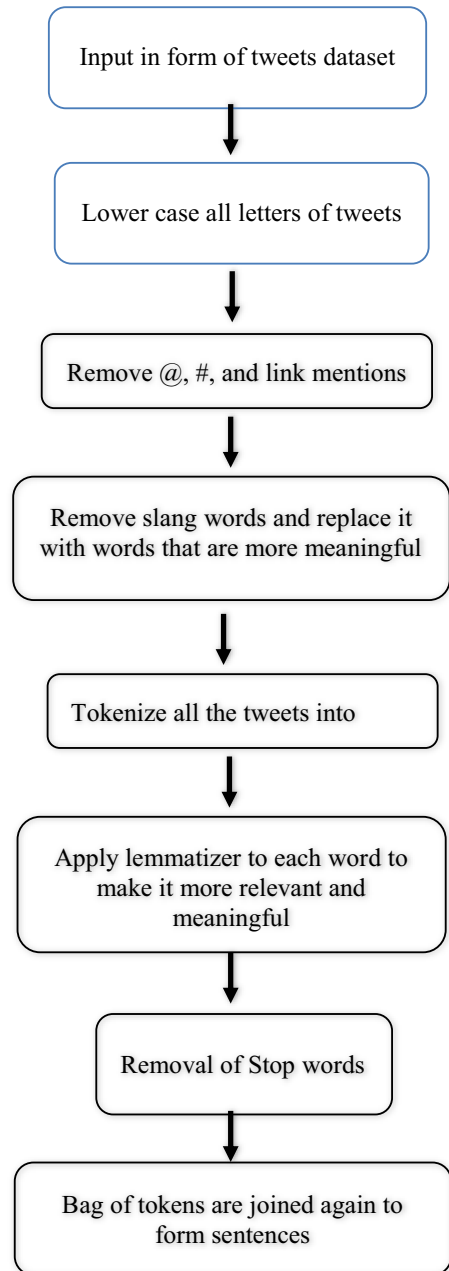


Fig. 2 Work flow diagram of proposed age and gender classification methodology

Fig. 3 Flow diagram for pre-processing



gave Tf-Idf score to these features, rather than to individual words. This produced better results than simply using the word unigrams. The expression for Tf-Idf is provided in eq. 1.

$$Tf-Idf = Term\ frequency * Inverse\ document\ frequency \quad (1)$$

where, Term Frequency = no of times a keyword appears in a document and Inverse document frequency (IDF) is defined as in eq. 2.

$$IDF = \log \left(1 + \frac{\text{Total number of documents}}{\text{Number of documents with keyword}} \right) \quad (2)$$

Tf-Idf is a better way of getting the most relevant words and characters from a corpus. It is better than counting the frequency of the terms because when considering only the frequency of terms, some common terms like ‘the’, ‘is’, ‘are’ etc. occur in the sentences more frequently than the others. These variables tend to overshadow the results produced by the classifier. Tf-Idf gives a lower score value to these terms and hence let the more relevant terms decide the output of the classifier.

N-grams are a contiguous sequence of n words from a text, tweet in this case. For example, “This is a tweet”, in this sentence unigrams are: *this, is, a, tweet*. Bigrams are: *this is, a tweet* etc. Now that text information is converted into word sequence, we need to convert word sequence into numerical features. Vectorization techniques are used for converting word sequences to numerical features. Some popular vectorization techniques are Binary Term Frequency, Bag of Words (BoW) Term Frequency, Normalized Term Frequency, Normalized Tf-Idf.

We used the sklearn’s TfidfVectorizer with an n-gram range of (1, 3) for words and (2, 5) for the characters, then used sklearn’s Pipeline to integrate the wordVectorizer and charVectorizer to form an integrated n-grams vectorizer which were used as the feature set.

4.3 Classification

Now that feature extraction phase is complete, we need to apply machine learning classifiers. The tweets are required to be classified into two classes for gender prediction: Female, Male and into four classes for age prediction: 18–24, 25–35, 35–49, and 50+ .

There are several well-known classification algorithms, which were used for sentiment analysis in the past. Classifiers such as logistic regression, SVM, Naive Bayes, decision tree and balanced winnow 2 were used the most in NLP [31]. Numerous studies opted for different classification algorithms and produced different results. In this study, we have used four different classifiers, discussed briefly as below:

4.3.1 Logistic regression

Logistic regression is a classification algorithm, which is very useful in binary classification i.e. whether the gender of the author is male or female. Unlike linear regression, it models the data using a non-linear function like the sigmoid function It can also be used for classification problems, where the number of classes in the output are greater than 2. The mathematical expression for sigmoid function is given in eq. 3.

$$\sigma_{sig}(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

4.3.2 Decision tree

A Decision Tree consists of two types of nodes, decision nodes (internal nodes) and leaf nodes. The top node of the tree is called the root node. The input parameters make up the internal

nodes and are used to make decisions and split the data further. Which input parameter will be used for the next node is decided by calculating the Gini impurity index. The parameter with the minimum Gini value is used for data splitting. Various algorithms like CART, ID3, C4.5 are also used to create decision trees. The mathematical expression for Gini impurity index is given as in eq. 4.

$$I_{Gini}(j) = \sum_i P(i|j)(1-P(i|j)) \quad (4)$$

4.3.3 Support vector machines (SVM)

The main idea behind an SVM classifier is to choose a hyperplane that can segregate n-dimensional data into different classes with minimum overlapping. Support vectors are used to create the hyperplane, hence the name ‘Support vector machines’. In an SVM model, the distance between a point x and the hyperplane, represented by (w, b) where,

$$||w|| = 1 \text{ is } |< w, x > + b| \quad (5)$$

4.3.4 Random Forest

Random forest classifiers are a part of ensemble-based learning methods. Their main features are Ease of implementation, efficiency and great output in a variety of domains. In the Random forest approach, many decision trees are constructed during the training stage. Then, a majority voting method is used among those decision trees during the classification stage to get the final output.

The model should be effective and unbiased for unseen data. Sometimes the dataset may contain redundant values in training data and testing data, and the model trained using the redundant data will just repeat the labels of the data that it has just seen during the training would have a perfect score but cannot predict anything accurately for unseen data. This situation is called overfitting. To avoid this situation and test the model for real problems, we applied k-fold cross-validation. Cross-validation is a technique used to test the effectiveness of a machine learning model. It is also a resampling procedure used to evaluate a model if we have limited data. In this technique, the data is split into k partitions, then train the model using the $k-1$ fold and validate the model using the k th fold and repeat this process until every partition serves as k th fold. Then take the average of the accuracies obtained in the above steps. 10-fold cross-validation was applied to our model and a significant change in the performance of the model was observed.

5 Implementation and result analysis

In this section, the implementation details, performance metrics used to evaluate the model followed by the results obtained are discussed in detail.

5.1 Implementation details

For pre-processing, Tweet cleaning is done via the pre-processor library. Nltk’s WordNetLemmatizer is used for lemmatizing the tokens in the tweets. Gensim library is used to remove the stop words from the lemmatized tweets.

sklearn’s TfidfVectorizer is used to create a numerical representation of the data and the dimensionality reduction is done using sklearn’s TruncatedSVD model.

For classification, we have used sklearn libraries classifiers. For this work, we used sklearn’s RandomForestClassifier, LinearSVC, DecisionTreeClassifier and LogisticRegression. Sklearn provides very efficient implementations of these algorithms. They are also provided in separate models which makes them easy to integrate into a new model.

5.2 Evaluation metrics

The most important step in the implementation of the proposed work is the evaluation of its performance. As a result, the evaluation of the machine learning algorithms used in this research is also needed. For this purpose, various types of measurement metrics are available. The metrics used in this article are discussed below.

5.2.1 Confusion matrix

The confusion matrix is a technique for the evaluation of the performance of a machine-learning model. The matrix compares the predicted values with actual target values. This provides us with a comprehensive picture of how well our classification model is doing and the types of errors it makes as shown in Fig. 4.

True Positive (TP): The model predicted positive and it is true.

False Positive (FP): The model predicted positive but it is false.

False Negative (FN): The model predicted negative and it is false.

True Negative (TN): The model predicted negative and it is true.

5.2.2 Accuracy

The percentage of accurate predictions for the test data is known as accuracy. It is figured out by dividing the number of valid predictions by the total number of predictions.

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{6}$$

Fig. 4 Confusion Matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

5.2.3 Precision

The number of positive class predictions that actually belong to the positive class is calculated by.

precision.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

5.2.4 Recall

The number of positive class predictions made out of all positive examples in the dataset is measured by the recall.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

5.2.5 F-measure

The harmonic mean of the two fractions, precision and recall is called F-measure. The F-measure attempts to strike a balance between precision and recall. The consequence is a number between 0.0 and 1.0, with 0.0 representing the worst F-measure and 1.0 representing the best F-measure.

$$F\text{-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

5.3 Result analysis

After observing from the literature review, the most efficient classifiers for such tasks were support vector machine, decision tree and random forest. In the beginning, we aimed to get results from the given dataset (PAN at CLEF 2015) and we have implemented the portion of the dataset, which is in the English language. Applied different features extraction methods to get the desired results. Afterwards, different classifiers such as SVM, logistic regression, decision tree and random forest. SVM gave us the best results. Then changed the parameters of the classifier to get more accurate but not much progress was observed. We were left with the last option i.e. making some tweaks in the dataset. So, first, we combined the dataset into 1 and then split it into 2 parts in a ratio of 3:7 (test dataset: training dataset) in order to increase the size of the training dataset and the result increased.

Table 5 and Table 6 lists results produced by using different classifiers for gender and age prediction. The decision tree classifier gave an accuracy of 67.7% whereas random forest was able to give 83.0% accuracy but the support vector machine (SVM) produced the best accuracy 96.6% for predicting the gender of tweets of unknown users.

Decision tree classifier gave an accuracy of 55.90% whereas random forest was able to give 69.49% accuracy but support vector machine (SVM) produced best accuracy of 83.05% for predicting age of tweets of unknown users.

Table 5 Performance metric for ‘Gender’ prediction

Performance Metrics	SVM	RF	DT	LR
Accuracy	0.933	0.831	0.678	0.8
F-measure	0.934	0.829	0.677	0.8
Precision	0.943	0.836	0.681	0.807
Recall	0.933	0.831	0.677	0.8

Table 6 Performance metric for ‘Age’ prediction

Performance Metrics	SVM	RF	DT	LR
Accuracy	0.833	0.694	0.559	0.8
F-measure	0.876	0.686	0.574	0.879
Precision	0.849	0.746	0.601	0.816
Recall	0.833	0.694	0.559	0.8

For better performance of the model and predict accurately for unseen data we must cross-check that the model is not over fitted. Overfitting occurs when a model would just repeat the labels of data that it has just seen while training. Such a model would have a perfect score but will fail to predict anything useful on unseen data. In this technique, we split data into k partitions, then train the model using the $k-1$ fold and validate the model using the k th fold. We repeat this process until every partition serves as a k th fold. Then take the average of the accuracies obtained in the above steps. We applied 10-fold cross-validation to the model. Table 7 and Table 8 shows the performance metric produced after applying 10-fold cross-validation.

After applying 10-fold cross-validation the performance of the model changed significantly. Accuracy for gender prediction came out to be 90.1% using SVM, 80.2% using Random forest and 66.9% using a decision tree. Similarly, accuracy for age prediction came out to be 80.2% using SVM, 67.6% using random forest and 54.4% using a decision tree. Table 9 and Table 10 shows performance metrics after applying 10-fold cross-validation with shuffling of data.

In section 2, the methodologies, features extraction methods and results of previous studies are listed in Table 2. The EG algorithm used in Argamon et al. [1] could give 80% accuracy. Whereas, Burger et al. [4] could give a maximum accuracy of 74% using Balanced-Winnow. Daneshvar et al. [7] used different classifiers and could achieve 81.8% maximum accuracy. Alowbidi et al. [8] were accurate to 78.3% using an NB-DT hybrid classifier and 5-g features. Using CNN Sezerer et al. [9] performed 75.1%. Sezerer et al. [10] applied CNN and RNN and achieved 78.47% accuracy using CNN and 82.31% using RNN. Rao et al. [27] used N-grams

Table 7 Performance metric for ‘Gender’ prediction after applying 10-Fold cross-validation

Performance Metric	SVM	RF	DT	LR
Accuracy	0.901	0.788	0.669	0.826
F-measure	0.901	0.788	0.668	0.827
Precision	0.907	0.799	0.683	0.834
Recall	0.901	0.788	0.669	0.826

Table 8 Performance metric for ‘Age’ prediction after applying 10-Fold cross-validation

Performance Metrics	SVM	RF	DT	LR
Accuracy	0.803	0.690	0.544	0.714
F-measure	0.809	0.765	0.556	0.775
Precision	0.820	0.722	0.588	0.739
Recall	0.803	0.690	0.544	0.714

Table 9 Performance metric for gender prediction after applying 10-Fold cross-validation with data shuffling

Performance Metrics	SVM	RF	DT	LR
Accuracy	0.880	0.798	0.632	0.812
F-measure	0.880	0.797	0.629	0.812
Precision	0.884	0.809	0.641	0.820
Recall	0.880	0.798	0.632	0.812

Table 10 Performance metric for age prediction after applying 10-Fold cross-validation with data shuffling

Performance Metrics	SVM	RF	DT	LR
Accuracy	0.810	0.681	0.537	0.725
F-measure	0.812	0.765	0.544	0.762
Precision	0.844	0.708	0.582	0.766
Recall	0.810	0.681	0.537	0.725

and stacked models along with SVM and scored an accuracy of 74.11% for age prediction and 72.33% for gender prediction.

As we can observe from the performance shown in Tables 11 and Fig. 5, our model outperformed the results of previous studies conducted. Our approach was able to give accuracy of 93.3% for gender prediction and 83.3% for age prediction without applying K-fold cross-validation and 88.0% accuracy for predicting gender and 81.0% accuracy for predicting age after applying 10-fold cross-validation.

Table 11 Comparison with some state-of-the-art methods

Article	Year	Dataset	method	Accuracy
Burger et al. [4]	2011	Twitter Dataset	Balanced Winnow, SVM, NB	74%
Daneshvar et al. [7]	2018	PAN 2018	NB, SVM, LR	81.8%
Alowbidi et al. [8]	2013	Twitter Dataset	NB, DT, NB-DT Hybrid	78.3%
Sezerer et al. [9]	2018	PAN 2018	CNN	75.1%
Sezerer et al. [10]	2018	PAN 2018	CNN and RNN	82.31%
Proposed method	2021	PAN-2015	SVM, Logistic Regression, Decision Tree and Random Forest	Gender: 88% Age: 81%

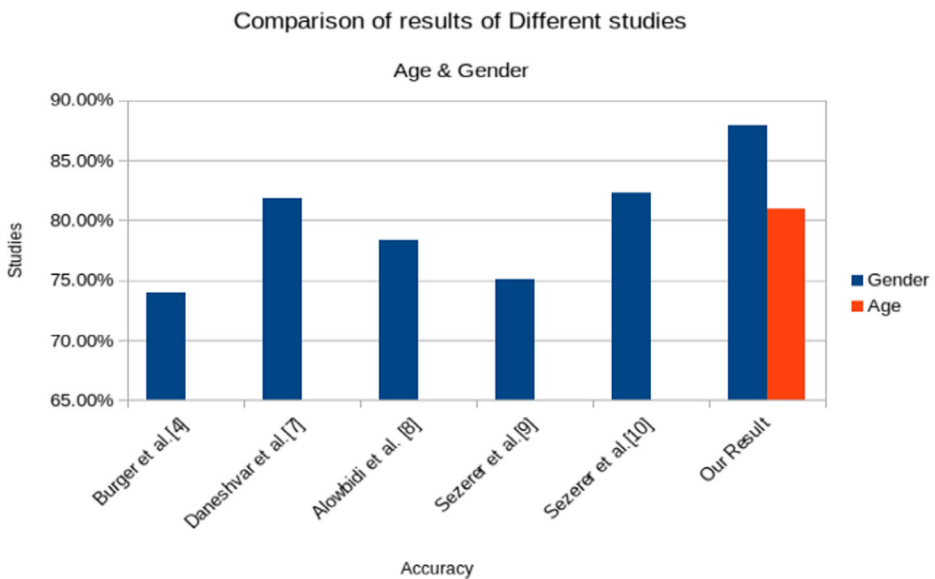


Fig. 5 Comparison of results with previous studies

6 Conclusion

In this work, we experimented with various feature extraction methods and classifiers to predict age and gender of a tweet using PAN at CLEF 2015 dataset. This intends to help in various fields such as marketing, personalization of user experience, legal investigations etc. We reviewed various previous researches and attempted to improve the accuracy for this project. NLP and machine-learning approach is proposed for author profiling. NLP techniques like tokenization, lemmatization, word and char n-grams are used in integration with machine learning classifiers like logistic regression, random forest, decision tree and SVM. So far, the SVM classifier was found to outperform other classification algorithms in age and gender prediction with an accuracy of 81.0% and 88.0% respectively.

In future, the work can further be extended in the following directions:

- One can contribute to the classification algorithm to improve the accuracy further.
- Incorporate the sentiments of the users for further study of their likes and dislikes and behaviour patterns.
- One can further extend the work to detect the criminals and self-harming activities such as suicide etc.
- One can include visual features such as photos and videos uploaded by the users for study.
- Cybersecurity aspects can also be considered to improve the model.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. *Text & Talk* 23(3):321–346
2. Herring SC, Paolillo JC (2006) Gender and genre variation in weblogs. *J Socioling* 10(4):439–459. <https://doi.org/10.1111/j.1467-9841.2006.00287.x>
3. Peersman C, Daelemans W, Van Vaerenbergh L (2011) Predicting age and gender in online social networks. In: *Proceedings of the 3rd international workshop on search and mining user-generated contents*, pp 37–44. <https://doi.org/10.1145/2065023.2065035>
4. Burger JD, Henderson J, Kim G, Zarrella G (2011) Discriminating gender on Twitter. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp 1301–1309
5. Ljubešić N, Fišer D, Erjavec T (2017) Language-independent gender prediction on twitter. In: *Proceedings of the second workshop on NLP and computational social science*, pp 1–6. <https://doi.org/10.18653/v1/W17-2901>
6. Rangel F et al (2018) Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, pp 1–38
7. Daneshvar S, Inkpen D (2018) Gender identification in twitter using n-grams and lsa. In: *proceedings of the ninth international conference of the CLEF association (CLEF 2018)*
8. Alowibdi JS, Buy UA, Yu P (2013) Empirical evaluation of profile characteristics for gender classification on twitter. In: *2013 12th international conference on machine learning and applications*, vol 1, pp 365–369
9. Sezerer E, Polatbilek O, Sevgili Ö, Tekir S (2018) Gender prediction from Tweets with convolutional neural networks: Notebook for PAN at CLEF 2018. In *19th working notes of CLEF conference and labs of the evaluation forum, CLEF 2018. CEUR Workshop Proceedings*
10. Sezerer E, Polatbilek O, Tekir S (2019) Gender prediction from tweets: improving neural representations with hand-crafted features. *arXiv preprint arXiv:1908.09919*
11. Arroju M, Hassan A, Farnadi G (2015) Age, gender and personality recognition using tweets in a multilingual setting. In *6th conference and labs of the evaluation forum (CLEF 2015): experimental IR meets multilinguality, multimodality, and interaction*, pp 23–31
12. Littlestone N (1988) Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Mach Learn* 2(4):285–318
13. Kivinen J, Warmuth MK (1997) Exponentiated gradient versus gradient descent for linear predictors. *Inf Comput* 132(1):1–63
14. Mendenhall TC (1887) The characteristic curves of composition. *Science* 9(214):237–249
15. Nguyen D, Smith NA, Rose C (2011) Author age prediction from text using linear regression. In *proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pp 115–123
16. Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B (2016) Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. *Working Notes Papers of the CLEF 2016*:750–784
17. Herring SC, Scheidt LA, Bonus S, Wright E (2004) Bridging the gap: A genre analysis of weblogs. In: *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 11. IEEE
18. Aries EJ, Johnson FL (1983) Close friendship in adulthood: conversational content between same-sex friends. *Sex Roles* 9(12):1183–1196
19. Jennifer C (2015) *Women, men and language 3rd edition*. Routledge, London. <https://doi.org/10.4324/9781315645612>
20. Heilbrun C (1988) *Writing a Woman's life*. Ballantine Books, New York
21. Tillery D (2005) The plain style in the seventeenth century: gender and the history of scientific discourse. *Journal of Technical Writing and Communication (JTWC)* 35(3):273–289. <https://doi.org/10.2190/MRQQ-K2U6-LTQU-0X56>
22. Savicki V, Lingenfelter D, Kelley M (2006) Gender language style and group composition in internet discussion groups. *J Comput-Mediat Commun* 2:0–0
23. Cieri C, Miller D, Walker K (2004) The fisher corpus: a resource for the next generations of speech-to-text. In: *LRE* 4:69–71
24. Pennebaker JW, Francis ME, Booth RJ (2001) *Linguistic inquiry and word count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates 71(2001)
25. DATASET: PAN at CLEF 2015: Author Profiling task <https://pan.webis.de/clef15/pan15-web/author-profiling.html>. Accessed on: 20 Aug 2020
26. Rexha A, Kröll M, Ziak H, Kem R (2018) Authorship identification of documents with high content similarity. *Scientometrics* 115(1):223–237
27. Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in twitter. In: *Proceedings of the 2nd international workshop on search and mining user-generated contents*, pp 37–44

28. Alowibdi JS, Buy UA, Yu P (2013) Language independent gender classification on twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 739–743
29. Okuno S, Asai H, Yamana H (2014) A challenge of authorship identification for ten-thousand-scale microblog users. In: 2014 IEEE international conference on big data (big data) (pp. 52–54). IEEE
30. Yadav D, Gupta A, Asati S, Choudhary N, Yadav AK (2020) Age group prediction on textual data using sentiment analysis. In: 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, pp 61–65
31. Bonaccorso G (2017) Machine learning algorithms. Packt Publishing Ltd
32. Chaffey D, Smith PR (2017) Digital marketing excellence: planning, optimizing and integrating online marketing. Routledge

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.