Journal on Big Data

Tech Science Press

DOI: 10.32604/jbd.2022.044060

ARTICLE

Author's Age and Gender Prediction on Hotel Review Using Machine Learning Techniques

Muhammad Hood Khan[1], **Bilal Khan**[1,*], **Saifullah Jan**[1] **and Muhammad Imran Chughtai**[2]

[1]Department of Computer Science, City University of Science and Information Technology, Peshawar, 25000, Pakistan

[2]Department of Computer Science, Sarhad University of Science and Information Technology, Mardan Campus, Mardan, Pakistan

*Corresponding Author: Bilal Khan. Email: bilalsoft63@gmail.com

Received: 20 July 2023 Accepted: 14 September 2023 Published: 17 November 2023

ABSTRACT

Author's Profile (AP) may only be displayed as an article, similar to text collection of material, and must differentiate between gender, age, education, occupation, local language, and relative personality traits. In several information-related fields, including security, forensics, and marketing, and medicine, AP prediction is a significant issue. For instance, it is important to comprehend who wrote the harassing communication. In essence, from a marketing perspective, businesses will get to know one another through examining items and websites on the internet. Accordingly, they will direct their efforts towards a certain gender or age restriction based on the kind of individuals who comment on their products. Recently many approaches have been presented many techniques to automatically detect user age and gender from the language which is based on text, documents, or comments on social media. The purpose of this research is to classify age (18–24, 25–34, 35–49, 50–64, and 65–70) and gender (male, female) from a PAN 2014 Hotel Reviews dataset of the English language. The usage of six machine learning models is the main emphasis of this work, including the methods of Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and K-Nearest Neighbors (KNN).

KEYWORDS

Author profiling; PAN-2014; machine learning; stylistic features

1 Introduction

The study of a specific collection of texts in an attempt to discover various characteristics of the author based on stylistic and content aspects, or to identify the author, is known as author profiling (AP). AP was used to be confined to hard-form documents, like newspaper articles and books. AP was used to identify and analyze multiple combinations of authors' textual characteristics, such as syntactic and lexical features. Content and stylistic features are the most useful characteristics for author profiling in digital writing [1]. Early research in author profiling mostly focused on a particular genre up to the change to author profiles on the Internet and social media. Digital text author profiling focuses on cross-genre author profiling, which employs one genre for training data and another for testing data, however, both must be relatively similar for successful results [2]. On social media, it is very easy to give a fake name, age, gender, and region to hide one's true character, but criminal activities

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

include, for example, pedophiles that have additional opportunities to connect with victims. Both law enforcement organisations and social network moderators are focusing on these issues in order to detect Internet predators [3].

Manual analysis virtually impossible on a large number of profiles and communications on social networks makes. To contact their victims, social media hunters frequently use false identities. Consequently, identifying and testing them is essential for a well-designed automated system. For example, anybody who wants to know the author's linguistic profile in the context of crime, security, or advertising has to be able to automatically retrieve information from text based on the author's gender, age, and other class characteristics. Aggressive text messages, or organizations seeking information about reviews of people who like or dislike their items, given sites and online item survey analytics resources [4]. Since data from social network sites, blogs and emails might all be mined online, the emergence of the internet in the twentieth and twenty-first centuries has catalyzed an increase in author profiling research. Web content has been analyzed in author profiling tasks to determine the age, gender, geographical history, country, and psychological characteristics of web users. A variety of applications, including forensics and marketing, have made use of the information acquired.

This study's main goal is to thoroughly investigate and assess a variety of known ML algorithms, together with cutting-edge feature extraction methods, within the specialized context of AP for age and gender prediction. The study makes use of the PAN 2014 Hotel Reviews dataset in an effort to identify complex stylistic patterns and characteristics that can successfully distinguish between age and gender categories in written material. Our study examines the performance of the SVM, KNN, RF, LR, NB, and DT models in a methodical manner, acknowledging the body of prior research on the subject while illuminating their application and effectiveness in this particular situation. We also explore the importance of aesthetic elements for textual data structure, giving a greater understanding of their influence on prediction accuracy. It is important to note that the contributions of our work include a thorough examination of a wide range of ML algorithms, the systematic assessment of feature extraction methods, and a thorough analysis of age and gender prediction within the provided dataset. Even if some earlier studies produced better outcomes, our research adds to the body of knowledge by offering a thorough comparison and highlighting the particular difficulties and possibilities this endeavor presents.

Rest of the paper is broken down into Section 2 concentrates on the related work. Section 3 presents the experimental setup and research methodology. Section 4 is a discussion on findings and its analysis. Finally, Section 5 concludes this study.

2 Related Work

This study focuses on the author's age and gender prediction. To this end, in the recent past many researchers have done work with their strengths and weaknesses. They have used various features and ML models in this regard where some of these features are: Semantic [5,6], Syntactic [2,7], Stylistic [3], Parts of speech (POS), and Lexical [8].

Researchers [4,9] used ReliefF, a feature selection method, to determine the gender of writers based on stylistic characteristics. The results showed that the ReliefF algorithm performed better in identifying important features than the original set of features, achieving high accuracy in gender prediction. In another study, the authors submitted their system based on 29 different stylistic features to the FIRE'18-MAPonSMS shared task, achieving an accuracy of 58.571% for the age and 73.714% for gender group on the training dataset. However, the accuracy dropped significantly on the testing data, with 0.37% and 0.55% for age and gender groups, respectively.

A system for determining age and gender categories from multilingual author profiles using machine learning techniques was proposed by [10]. The system uses an ensemble model composed of four classifiers to achieve high sensitivity and specificity for gender with an accuracy of 83%. However, the accuracy is relatively low for the age category at 60% and joint age and gender category at 49%. The text source can be automatically categorized and diagnosed with unknown testing data.

To determine the best ML technique for gender attribution in Russian language texts while removing topic and genre-related biases, [11] used a dataset of 1,456 texts written by 732 male and 724 female authors to extract topic-independent features. SVM and Random Forest were the most accurate algorithms in terms of accuracy and F1-score, according to the study's evaluation of several machine learning techniques. The study revealed that gender attribution in text classification tasks can be successfully accomplished using machine learning approaches, and it also illustrated how eliminating topic and genre-related biases may enhance the models' fairness and accuracy.

Authors in [7] suggested a content-based approach for determining the gender and age of writers with comparable writing styles. To train classifiers for various profile sizes, they utilise a number of features, including syntactic n-grams, part-of-speech tags, word and character n-grams, and word and character similarity. Word uni-grams and character tri-grams are the standard strategies. For age group and gender, respectively, the combination of word n-grams of different sizes yields the maximum accuracy of 0.496% and 0.734%.

These two studies conducted by [8,12], focused on gender identification and bot detection on Twitter using machine learning techniques. In the first research, 10 different classifiers' performance on 7 CLEFPAN collections is compared, and a 2-stage feature selection method is suggested to minimise feature size without noticeably lowering performance. The greatest results are often obtained using neural networks or random forest models, and the feature set size can be decreased to about 300 words. The second study explores semantic, syntactic, and stylistic features to engineer the feature set, combining them with part of speech tags. AdaBoost is the algorithm that achieves the highest F1-score of 0.99% on the development set for bot detection when they apply ensemble approaches. With this method, they are able to detect bots in English-language tweets with an accuracy score of 89.17%.

Martinc et al. [13] focused on author profiling, specifically gender and language variety prediction in tweets. The proposed approach includes tweet pre-processing, feature construction, feature weighting, and classification model construction. The authors used a Logistic regression classifier with various characters and word n-grams as main features. The results showed the best accuracy on the Portuguese test set for both gender and language variety prediction, with 0.8600% and 0.9838%, respectively. However, the worst accuracy was obtained on the Arabic test set.

Machine learning-based approaches for author profiling, specifically gender and age prediction were used by [14,15]. A new model was proposed by Reddy et al., which calculates document weights using a combination of POS N-grams and most frequent terms, achieving promising results in predicting gender from reviews. To predict gender and age from tweets, Katna et al. used ML classifiers such as logistic regression, decision tree, random forest, and support vector machines as well as Natural Language Processing (NLP) techniques like lemmatization, tokenization, word and character n-grams. The SVM classifier fared better than the others, obtaining accuracy levels of 88.0% for predicting gender and 81.0% for predicting age.

A statistical approach to author profiles on social media platforms based on the extraction of 17 stylometry-based features from users' tweets, was proposed by [16]. The authors used the technique of random forests to train their model and evaluate its performance on bot detection and gender classification tasks for both English and Spanish languages. The approach achieved the best performance with accuracy, precision, recall, and F1-score measures. The results showed that the proposed approach obtained an accuracy of 92.45% and 90.36% for bot detection and gender classification, respectively, in the English dataset, and an accuracy of 89.68% and 88.88% in the Spanish dataset.

To improve author profiling and performance by estimating the gender and age of users based on their textual productions. To achieve the best classification, the researchers [17] used a variety of ML algorithms such as RF, SVM, MP, DT, NB, LSTM and KNN on an English corpus taken from the PAN-AP-2015 dataset obtained from Twitter. The results showed that the effectiveness of each technique varied depending on the dataset. Deep learning techniques were found to be helpful when the dataset was large.

A study by Siddique at al., focused on gender identification and bot detection among English-speaking Twitter users [18]. Their method, which combined a bag of words model with a variety of pre-processing approaches, produced noteworthy results. They used Logistic Regression to get an accuracy of 87.12% for Task-A, which involves bot detection, while utilising a Decision Tree classifier to achieve an accuracy of 68.99% for Task-B, which involves gender identification. During the TIRA phase, which was used to assess their performance further, they achieved 68.37% accuracy for Task-B and 86.29% accuracy for Task-A. This investigation, carried out as a component of PAN at CLEF 2019, offers important new perspectives on the complex fields of bot identification and gender profiling on the Twitter network.

Ashraf et al. [19] proposed a benchmark corpus of bilingual (English and Roman-Urdu) tweets with an emphasis on author profiling on bilingual data. The corpus includes 339 AP annotated with six different traits. The study applies a range of deep learning methods, including CNN, LSTM, Bi-LSTM, and GRU, to the proposed corpus for age and gender identification. The results showed that the best performance was achieved by the Bi-LSTM method for age identification and gender identification tasks. The accuracy and F1-measure for gender identification are 0.882% and 0.839%, respectively, while for age identification, the accuracy and F1-measure are 0.735% and 0.739%, respectively.

Ouni et al. proposed two models to solve the author profiling problem, which aims to differentiate between humans and bots on Twitter and identify the gender of users [20]. The first model is topic-based and uses semantic and stylistic features to extract information from English tweets, which are integrated into a convolutional neural network. The 2nd model is classification model used on a Spanish corpus that makes use of statistical features to create a random forest-based classifier. The proposed models were evaluated on various standard databases and were found to be effective in terms of accuracy, G-mean, F1-score, recall, and precision. Overall summary of the literature is presented in Table 1.

Table 1: Summary of the related work

Study	Methods used	Results
[4,9]	ReliefF for stylistic gender prediction	High accuracy in gender prediction
[10]	Ensemble model for age and gender prediction	High sensitivity and specificity for gender
[11]	SVM, random forest for gender attribution	Successful gender attribution, bias reduction

(Continued)

Table 1 (continued)

Study	Methods used	Results
[7]	Content-based approach for gender and age	Accuracy: 0.496% (age), 0.734% (gender)
[8,12]	Various classifiers for gender identification	Accuracy up to 89.17% for bot detection
[13]	Logistic regression for gender, language	Accuracy: 0.8600% (gender), 0.9838% (language)
[14,15]	ML classifiers, NLP for gender and age	Accuracy: 88.0% (gender), 81.0% (age)
[16]	RF for bot, gender classification	Accuracy: 92.45% (bot), 90.36% (gender)
[17]	RF, SVM, LSTM, KNN for gender and age	Varying effectiveness based on dataset
[18]	Bag of words, LR, DT	Accuracy: 87.12% (bot detection), 68.99% (gender)
[19]	Deep learning methods for age, gender	Accuracy: 0.735% (age), 0.882% (gender)
[20]	Topic-based CNN, statistical model for bot, gender	Effective accuracy, G-mean, F1-score, recall, precision

3 Design of the Experiment and Research Methodology

This work emphases on APs using gender and age on PAN 2014, hotel reviews dataset written in the English language. All experiments are done on the system with the specification of Processor Intel (R) Core (TM) i5-10210U, RAM 8 GB, Window 11, Version 22H2, OS build 25145.1000, and architecture 64 bit. Tools used for research implementation are Anaconda Navigator, Jupyter Notebook, Spyder, and PAN 2014 Hotel Reviews dataset. The overall methodology followed in this study is presented in Fig. 1. The methodology used to classify authors by age and gender includes numerous pre-processing steps, including feature extraction from the data, and data pre-processing methods. These procedures are intended to convert unstructured text data into inputs for ML models that may be used to successfully learn from.

3.1 Description and Preprocessing of the Dataset

All tests are performed on the PAN 2014 Hotel Reviews dataset[1] available in the English Language. Gender is classified into two categories male and female while age is classified into five categories that are 18–24, 25–34, 35–49, 50–64, and 65–70 years, respectively. As the dataset is in textual form, some pre-processing steps are applied to structure the data, so that ML models can be easily trained. The Natural Language Processing (NLP) system's pipeline begins with text preprocessing, which might affect how well the system performs in the end [21]. Because real-world data might be sparse at times, preprocessing is the act of transforming raw data into a comprehensible format, inconsistent, redundant, and loud [22,23]. The PAN 2014 Hotel reviews dataset was used for age and gender classification due to its consistency with the study aims and real-world applicability. The dataset contains genuine, different hotel evaluations with varying writing styles, making it useful for extracting stylistic elements. The use of this data is clear in its applications, such as marketing, where organisations analyse reviews to adjust plans, and security, where author profiling assists in the identification of

[1] https://www.kaggle.com/datasets/datafiniti/hotel-reviews.

risks. Furthermore, the dataset reflects current communication settings by mirroring online social interactions. Because of this, the dataset is an excellent candidate for investigating the accuracy and consequences of age and gender categorization using machine learning models.

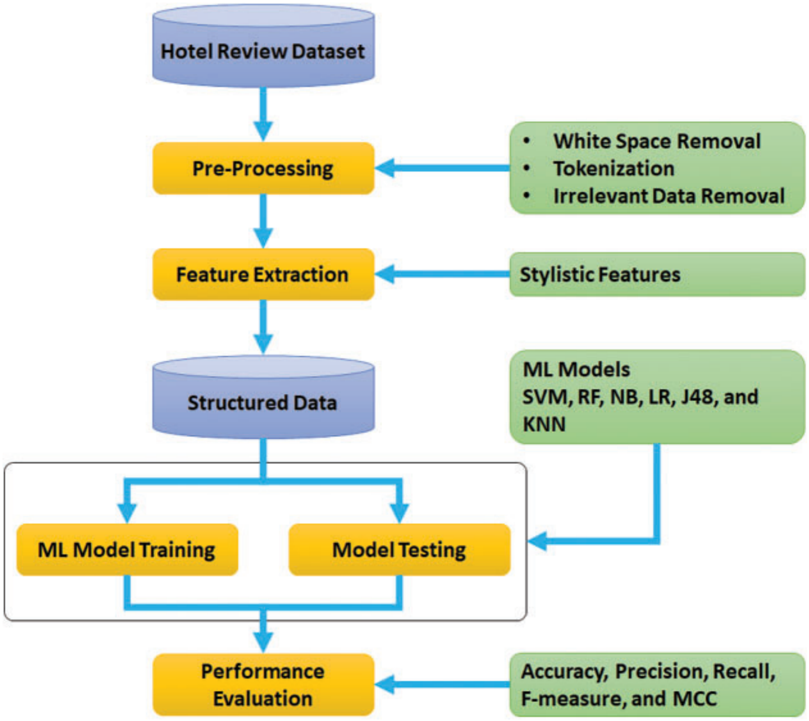

Figure 1: Our research approach

Through pre-processing procedures removed some irrelevant data from the dataset, such as @replies, hashtags, and URL links, thus how to appropriately handle this bias is a crucial issue. This study focused on some preprocessing approaches which are white space removal, tokenization, and irrelevant data removal. An example of each of them is shown in Fig. 2.

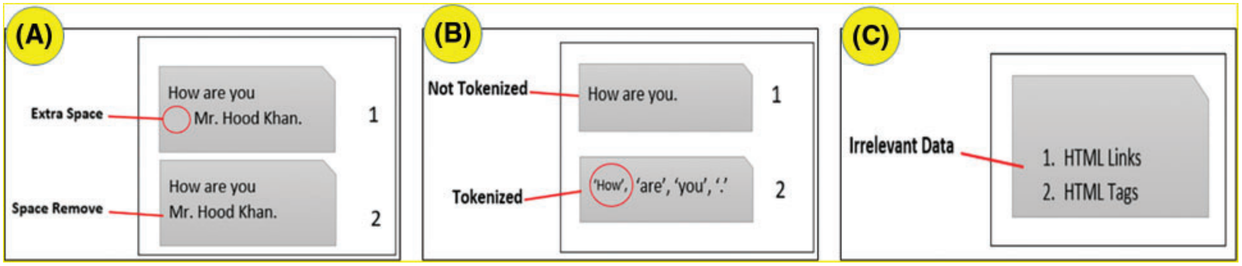

Figure 2: (A) White space removal, (B) Tokenization, and (C) Irrelevant data removal

3.2 Features Extraction

Feature extraction refers to the process of selecting or extracting relevant information or features from a dataset, which can be used as input to the ML models or other analyses. The goal is to reduce the amount of data to a more manageable size while retaining important information that can be used for further analysis. Feature extraction is commonly used in image processing, natural language processing, and other fields where large datasets need to be analyzed [24].

Text data must be converted into numerical representations that ML algorithms can understand, and this is where feature extraction comes in. As mentioned in the article, the aesthetic elements are the main focus here. Each stylistic element helps to identify a writer's writing style and may reveal information about the author's age and gender. Stylistic features refer to the distinctive features of a writer's style or a particular type of writing. These features can include elements such as tone, diction, syntax, imagery, figurative language, and others. They contribute to the unique voice, mood, and atmosphere of a piece of writing, and help to convey meaning in a creative and impactful way [25]. We have extracted 14 stylistic characteristics that are presented in Table 2.

Table 2: List of stylistic features

S. No.	Stylistic-feature	Description
1	Avg. length of word	It determines its average after counting the words in a text document.
2	Avg. length of sentence	It determines its average after counting all the sentences in a text document.
3	% of words with six or more letters	Percentage of words that only contain six letters or more.
4	% of words with two or three letters	Percentage of words with just two or three letters.
5	% of question sentences	A written text document's percentage of question sentences.
6	% age of semicolons	It refers to how many percentages of semicolons there are in a written document.
7	% age of punctuations	It determines the percentage by counting all the punctuations in a text document.
8	% age of comma	It determines the percentage after counting all the commas in sentences.
9	% age of short sentences	It determines the percentage of sentences that include four or fewer words after counting all of them.
10	% age of long sentences	It determines the percentage of sentences that include more than four words after counting all of them.
11	% age of capitals	It is described as calculating the percentage from the total number of capital letters used in text content.
12	% age of colons	In a text document, it is the percentage of colons that are utilised.

(Continued)

Table 2 (continued)

S. No.	Stylistic-feature	Description
13	% age of digits	It is described as the calculation of the percentage after counting all the digits in a text document.
14	% age of full stop	Total number of full-stop sentences used in a written work, which is then counted and the percentage calculated.

3.3 Models Training and Performance Evaluation

In this work, there are six well-known ML models are used on PAN 2014 dataset. These models include SVM [26,27], RF [28,29], LR [30,31], NB [32,33], KNN [29,34], and J48 [28,35]. These models are trained using 10-fold cross-validation methods. The selection of models like is basedon their well-established status as benchmark algorithms in ML, allowing for comparative analysis and insights into model performance across the task of age and gender prediction from text. Additionally, factors includes model fit, approach variety, library accessibility, interpretability, and prospective research-related discoveries may have played a role in the decision. Efficacy of each model is calculated using some standard evaluation measures including precision, recall, F-measurem Methew's Coorelation Coefficient (MCC), and accuracy.

4 Results Analysis and Discussion

This section presents the outcomes assessed through various measures. Overall measures are calculated from the confusion matrix (CM) of each model. CM is a widely used metric for categorization problems; it may be used to solve both binary and multiclass classification issues. Table 3 presents the CM of each employed model for age prediction. It is a multi-class problem having five classes that are 18–24, 25–34, 35–49, 50–64, and 65–70. Table 4 presents the CM values of each model for gender prediction which is a binary classification problem.

Table 3: Confusion matrix of each model for age prediction

S. No.	Techniques	Age	18–24	25–34	35–49	50–64	65–70
1	LR	18–24	0	1	4	4	0
		25–34	0	0	15	24	0
		35–49	0	0	16	25	0
		50–64	0	2	9	23	0
		65–70	0	0	4	23	0
2	RF	18–24	0	6	1	2	0
		25–34	0	5	9	15	10
		35–49	0	4	11	12	14
		50–64	0	6	11	11	6
		65–70	0	2	9	10	6

(Continued)

Table 3 (continued)

S. No.	Techniques	Age	18–24	25–34	35–49	50–64	65–70
3	NB	18–24	5	1	1	0	2
		25–34	19	1	5	0	14
		35–49	16	1	4	2	18
		50–64	15	2	4	1	12
		65–70	8	2	4	1	12
4	SVM	18–24	0	2	0	6	1
		25–34	0	2	6	29	2
		35–49	0	2	9	27	3
		50–64	0	2	4	24	4
		65–70	0	4	7	12	4
5	DT	18–24	0	1	2	3	3
		25–34	4	7	7	9	12
		35–49	3	8	8	13	9
		50–64	1	5	12	9	7
		65–70	2	1	6	14	4
6	KNN	18–24	1	1	3	2	3
		25–34	4	8	9	9	3
		35–49	3	6	13	8	8
		50–64	3	13	5	7	10
		65–70	3	3	11	8	6

Table 4: Confusion matrix of each model for gender prediction

S. No.	Techniques	Gender	Male	Female
1	LR	Male	78	4
		Female	65	3
2	RF	Male	52	30
		Female	33	35
3	NB	Male	78	4
		Female	64	0
4	SVM	Male	65	17
		Female	51	17
5	DT	Male	50	32
		Female	28	40
6	KNN	Male	19	52
		Female	35	44

Tables 5 and 6 present the precision, recall, and F1-score achieved through each model for age and gender prediction, respectively. The research demonstrates that SVM performs better for age prediction with values of 0.26, 0.209, and 0.225 for precision, recall, and F1-score accordingly. However, NB shows the worst performance with 0.18, 0.231, and 0.131 for precision, recall, and F1 for age prediction. Furthermore, for gender prediction, DT outperforms other employed models with values of 0.63, 0631, and 0.631 for precision, recall, and F1-score, respectively, while KNN shows the worst performance with values of 0.472, 0.475, and 0.464 accordingly for precision, recall, and F1-score.

Table 5: Performance analysis of each employed model for age prediction using precision, recall, and F1-score

S. No.	Techniques	Precision	Recall	F1-score
1	LR	0.11	0.207	0.138
2	RF	0.211	0.216	0.206
3	NB	0.18	0.231	0.131
4	SVM	0.209	0.225	0.181
5	DT	0.165	0.16	0.16
6	KNN	0.21	0.212	0.211

Table 6: Performance analysis of each employed model for gender prediction using precision, recall, and F1-score

S. No.	Techniques	Precision	Recall	F1-score
1	LR	0.487	0.497	0.386
2	RF	0.575	0.574	0.574
3	NB	0.267	0.476	0.342
4	SVM	0.53	0.521	0.495
5	DT	0.63	0.631	0.631
6	KNN	0.472	0.475	0.464

The MCC takes into account TRUE Negatives, TRUE Positives FALSE Negatives, and FALSE Positives, produces value between -1 and + 1, where + 1 indicates perfect agreement between predicted and observed classifications, 0 indicates no agreement beyond chance, and -1 indicates complete disagreement. Fig. 3 presents the MCC analysis of each employed model for age prediction. This study shows SVM's superior performance and the worst performance of DT for age prediction. On the other hand, Fig. 4 presents the MCC analysis of utilized models for gender prediction. In this case, DT outperforms the other models used, while KNN performs worst of all.

Accuracy analysis is important because it enables us to assess performance of a model or system in making correct predictions or classifications. By measuring accuracy, we can assess how effectively the model generalises to new, previously unknown data. Moreover, accuracy analysis helps in comparing different models and selecting the best-performing one for the given task. Overall, accuracy analysis is a fundamental step in assessing the effectiveness of a model or system and improving its performance.

Fig. 5 presents the accuracy analysis of each model for age prediction. Same as the previous analysis for gender prediction, accuracy also show SVM's superior performance with a success rate of 0.26, on the other side, for gender prediction, Fig. 6 illustrates DT's superior performance with a value of 0.63. The overall analysis summarises that for age prediction we can use SVM as compared with other employed models however, for gender analysis, we can focus on DT.

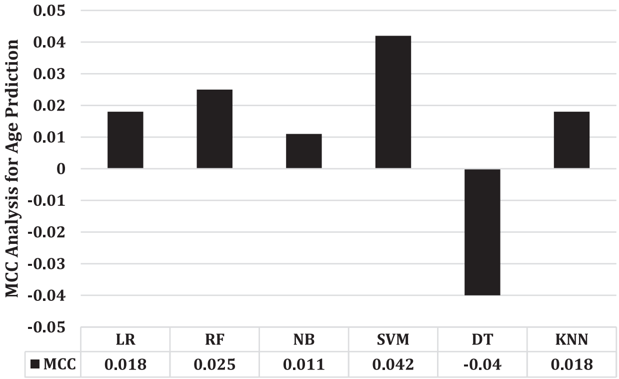

Figure 3: MCC performance analysis for age prediction

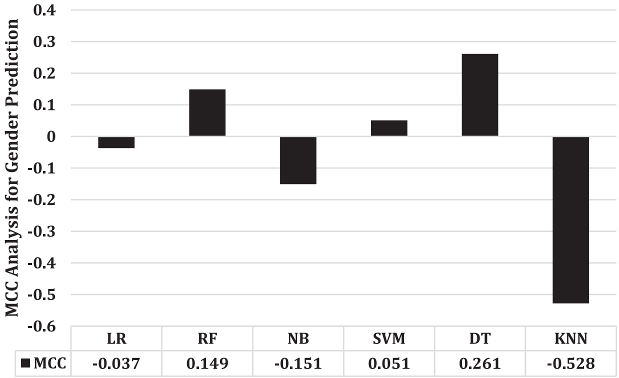

Figure 4: MCC performance analysis for gender prediction

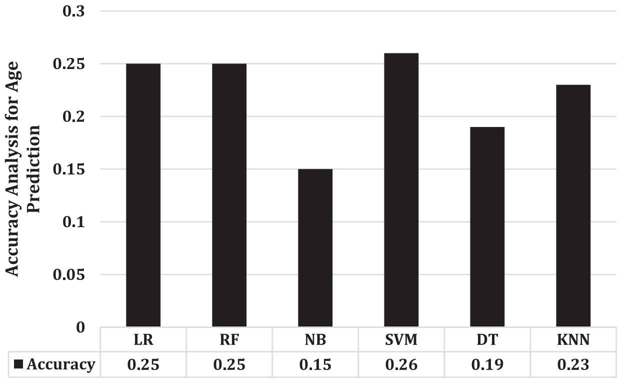

Figure 5: Accuracy analysis for age prediction using each employed model

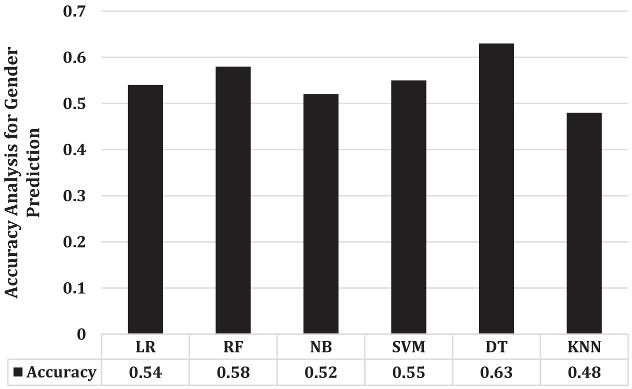

Figure 6: Accuracy analysis for gender prediction using each employed model

Now, for further analysis, there is a need to find the difference between the outcomes achieved. To do this, we are performing the percentage difference which is a way of expressing the difference between two values as a percentage of their average value. The absolute value of the percentage difference is taken to ensure that the result is always positive. A percentage difference of zero indicates that the two values are equal, while a percentage difference greater than zero indicates that one value is larger than the other. The percentage difference is often used in scientific and engineering contexts to compare experimental and theoretical values, as well as in financial analysis to compare changes in stock prices or other market indicators over time. It can be calculated as:

$$P\,D = ((n1 - n2)/(n1 + n2/2)) * 100 \tag{1}$$

Fig. 7 depicts the percentage difference between SVM and other applicable models. This research shows that there is only a 2.73% difference between the results of SVM with LR and RF. However, Fig. 8 illustrates that RF as compared with LR, DT, NB, and KNN, has the minimum difference from SVM.

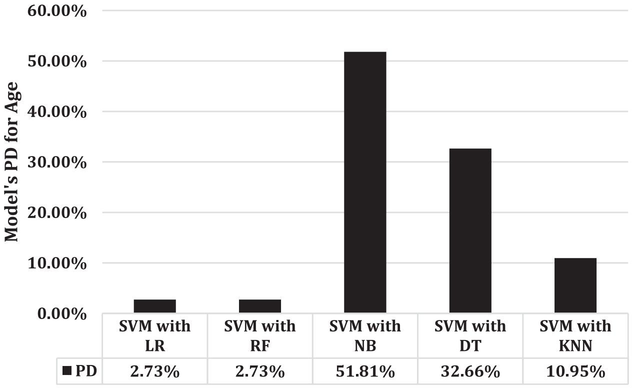

Figure 7: Percentage difference for age prediction using SVM compared to other models

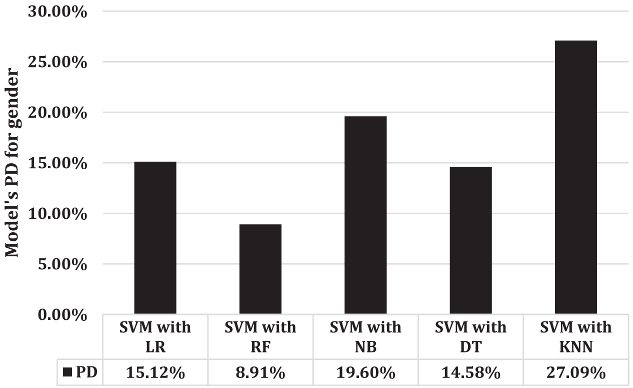

Figure 8: DT's gender prediction performance compared to other models, in percentage

In this study, we have utilized two types of classification problems which are binary class problems for gender prediction, and multi-class problems for age prediction. SVM performs better on multiclass datasets because it can efficiently handle high-dimensional feature spaces and non-linear decision boundaries. Additionally, SVMs can be trained with different types of kernel functions that can better capture the underlying structure of the data. This allows SVM to accurately classify data points into multiple classes, even when there is significant overlap between them. SVM also has a regularization parameter that helps prevent overfitting, making it a robust and effective algorithm for multiclass classification problems. DT works well on binary class datasets because they can naturally split the data into two groups based on the binary outcome. This allows the tree to make simple and easily interpretable decisions about which features are important in predicting the binary outcome. Additionally, DT can handle imbalanced data well, which is common in binary classification problems. However, DT may not perform as well on multi-class datasets because they may struggle to find optimal splits when there are more than two possible outcomes.

To improve the interpretability of the results and acquire understanding of the age and gender classification task described in the research, feature significance analysis must be conducted. Researchers can identify the essential characteristics underlying precise predictions by analysing the impact of linguistic elements, such as by permutation importance, tree-based model splits, Least Absolute Shrinkage and Selection Operator (LASSO) regression, or SHapley Additive exPlanations (SHAP) values. In the PAN 2014 Hotel Reviews dataset, our study clarifies the distinctive writing styles that distinguish age and gender groups, strengthening our grasp of underlying trends and enabling more intelligent model result interpretations.

There are a few limitations that should be taken into account even though the paper offers a thorough analysis of age and gender categorization using ML models on the PAN 2014 Hotel Reviews dataset. First of all, the study only considers stylistic characteristics for author profiling, thereby ignoring additional pertinent language clues that can help with precise predictions. The study also uses a predetermined set of ML models, thereby avoiding the investigation of more recent or sophisticated algorithms that may result in better results. Also notable is the lack of a thorough examination of the ethical ramifications of author profiling, particularly with regard to privacy and potential abuse. Last but not least, the study makes no mention of alternative methods for enhancing model resilience, such data augmentation or ensemble procedures.

5 Conclusion

AP for age and gender prediction is an NLP task that involves using computational techniques to predict the gender and age of an author based on their written text. It can be achieved using various ML models, as some of these used in this study are SVM, KNN, RF, LR, NB, and DT. For AP, this study focuses on the stylistic features to structure the data. However, the accuracy of the prediction may depend on various factors, such as the quality of the data, the diversity of the language used, and the cultural and linguistic norms of the target audience. The aim of this study is to investigate how well several ML models (SVM, RF, NB, LR, DT, and KNN) perform in properly detecting age and gender based on stylistic elements derived from a PAN 2014 Hotel reviews dataset. The study seeks to know how well the applied models can classify age and gender based on aesthetic elements derived from a PAN 2014 Hotel reviews dataset. What are the consequences of these classification results for numerous practical applications in security, forensics, marketing, and other information-related domains? The study also investigates the possible uses of such categorization in disciplines such as security, forensics, marketing, and other information-related fields. The results of this investigation can be used by firms to identify the age and gender of their customers and accordingly promote their products and services. Overall, the conclusions of this study can help to build automatic author profiling techniques, which can have important practical applications in different fields. The overall outcomes show better performances of SVM and DT for age and gender, respectively.

Future work on this study may use cutting-edge methods like multimodal analysis to combine textual and visual indicators for author profiling. The accuracy, morality, and flexibility of age and gender categorization algorithms might also be improved by addressing potential biases and assuring fairness in predictions, looking into privacy-preserving techniques, and researching cross-linguistic variances in writing styles. The dynamic nature of author traits may be better understood with the help of more study into real-time profiling, longitudinal analysis, and human-AI collaboration, which may also increase the practical application of the suggested methodologies in a variety of real-world contexts.

Acknowledgement: We gratefully acknowledge the collaborative efforts of all four authors in this research, who contributed equally and passionately to every aspect of the study. Our combined dedication and insights have enriched this work, reflecting the true essence of teamwork and shared commitment.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Muhammad Hood Khan (MH), Bilal Khan (BK); data collection: Saifullah Jan (SJ), Muhammad Imran Chughtai (MI); analysis and interpretation of results: MH, BK, SJ; draft manuscript preparation: MI. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All the data used in this study is online available at: https://www.kaggle.com/datasets/datafiniti/hotel-reviews.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. P. López-Monroy, M. Montes-y-Gómez, H. J. Escalante, L. Villasenor-Pineda and E. Stamatatos, "Discriminative subprofile-specific representations for author profiling in social media," *Knowledge-Based Systems*, vol. 89, pp. 134–147, 2015.

[2] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka and I. Moloshnikov, "Machine learning models of text categorization by author gender using topic-independent features," *Procedia Computer Science*, vol. 101, pp. 135–142, 2016.

[3] T. K. Koch, P. Romero and C. Stachl, "Age and gender in language, emoji, and emoticon usage in instant messages," *Computers in Human Behavior*, vol. 126, pp. 106990, 2022.

[4] J. Silva, S. García, M. A. Binda, F. M. Gonzalez, R. Barrios *et al.,* "A method for detecting the profile of an author," *Procedia Computer Science*, vol. 170, pp. 959–964, 2020.

[5] F. Pulvermüller, "Neurobiological mechanisms for semantic feature extraction and conceptual flexibility," *Topics in Cognitive Science*, vol. 10, no. 3, pp. 90–620, 2018.

[6] Q. Yuan, H. Z. Shafri, A. H. Alias and S. J. Hashim, "Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data," *Remote Sensing*, vol. 13, no. 13, pp. 2473, 2021.

[7] I. Ameer, G. Sidorov and R. M. A. Nawab, "Author profiling for age and gender using combinations of features of various types," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 5, pp. 4833–4843, 2019.

[8] G. Kovács, V. Balogh, P. Mehta, K. Shridhar, P. Alonso *et al.,* "Author profiling using semantic and syntactic features notebook for PAN," in *Conf. and Labs of the Evaluation Forum (CLEF), CEUR Workshop Proc.*, September, Lugano, Switzerland, vol. 2380, pp. 9–12, 2019.

[9] T. Raghunadha Reddy, B. Vishnu Vardhan, M. GopiChand and K. Karunakar, "Gender prediction in author profiling using ReliefF feature selection algorithm," in *Intelligent Engineering Informatics: Proc. of the 6th Int. Conf. on FICTA*, Bhubaneswar, Odisha, pp. 169–176, 2018.

[10] A. Sittar and I. Ameer, "Multi-lingual author profiling using stylistic features," in *Sun SITE Central Europe Workshop Proc.*, Gandhinagar, India, vol. 2266, pp. 240–246, 2018.

[11] A. Nemati, "Gender and age prediction multilingual author profiles based on comments," in *Sun SITE Central Europe Workshop Proc.*, Gandhinagar, India, vol. 2266, pp. 232–239, 2018.

[12] C. Ikae and J. Savoy, "Gender identification on Twitter," *Journal of the Association for Information Science and Technology*, vol. 73, no. 1, pp. 58–69, 2022.

[13] M. Martinc, I. Skrjanec, K. Zupan and S. Pollak, "PAN 2017: Author profiling-gender and language variety prediction," in *Conf. and Labs of the Evaluation Forum (Working Notes)*, Dublin, Ireland, 2017. https://api.semanticscholar.org/CorpusID:20540232

[14] T. R. Reddy, B. V. Vardhan and P. V. Reddy, "N-gram approach for gender prediction," in *Proc. of 7th IEEE Int. Advanced Computing Conf., IACC 2017*, Hederabad, India, pp. 860–865, 2017.

[15] R. Katna, K. Kalsi, S. Gupta, D. Yadav and A. K. Yadav, "Machine learning based approaches for age and gender prediction from tweets," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27799–27817, 2022.

[16] S. Ouni, F. Fkih and M. N. Omri, "Toward a new approach to author profiling based on the extraction of statistical features," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–16, 2021.

[17] R. Alroobaea, "An empirical combination of machine learning models to enhance author profiling performance," *International Journal*, vol. 9, no. 2, pp. 2130–2137, 2020.

[18] M. H. F. Siddiqui, I. Ameer, A. F. Gelbukh and G. Sidorov, "Bots and gender profiling on Twitter," in *Conf. and Labs of the Evaluation Forum (Working Notes)*, Kolkata, India, 2019. https://ceur-ws.org/Vol-2380/paper_186.pdf

[19] M. A. Ashraf, R. M. A. Nawab and F. Nie, "Author profiling on bi-lingual tweets," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 2379–2389, 2020.

[20] S. Ouni, F. Fkih and M. N. Omri, "Novel semantic and statistic features-based author profiling approach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12807–12823, 2023.

[21] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint*, arXiv:1707.017802017, 2018.

[22] A. Grivas, A. Krithara and G. Giannakopoulos, "Author profiling using stylometric and structural feature groupings," in *Notebook for PAN at Conf. and Labs of the Evaluation Forum*, Toulouse, France, 2015. https://api.semanticscholar.org/CorpusID:11729171

[23] T. Wichaiwong, K. Koonsanit and C. Jaruskulchai, "A simple approach to optimized text compression's performance," in *Proc. of Int. Conf. on Next Generation Web Services Practices (NWeSP)*, Seoul, Korea, pp. 66–70, 2008.

[24] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *International Scholarly Research Notices (ISRN) Neuroscience*, vol. 2014, no. 3, pp. 1–7, 2014.

[25] I. Pervaz, I. Ameer, A. Sittar and R. M. A. Nawab, "Identification of author personality traits using stylistic features: Notebook for PAN at CLEF 2015," in *Conf. and Labs of the Evaluation Forum (Working Notes)*, Toulouse, France, pp. 1–7, 2015.

[26] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang *et al.,* "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[27] M. Al-Qatf, Y. Lasheng, M. Al-Habib and K. Al-Sabahi, "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection," *IEEE Access*, vol. 6, pp. 52843–52856, 2018.

[28] R. Naseem, B. Khan, A. Ahmad, A. Almogren, S. Jabeen *et al.,* "Investigating tree family machine learning techniques for a predictive system to unveil software defects," *Complexity*, vol. 2020, pp. 1–21, 2020.

[29] R. Naseem, B. Khan, M. A. Shah, K. Wakil, A. Khan *et al.,* "Performance assessment of classification algorithms on early detection of liver syndrome," *Journal of Healthcare Engineering*, vol. 2020, pp. 1–13, 2020. https://doi.org/10.1155/2020/6680002

[30] B. Khan, R. Naseem, F. Muhammad, G. Abbas and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020.

[31] B. T. Pham, "A novel classifier based on composite hyper-cubes on iterated random projections for assessment of landslide susceptibility," *Journal of the Geological Society of India*, vol. 91, no. 3, pp. 355–362, 2018.

[32] A. Iqbal, S. Aftab, U. Ali, Z. Nawaz, L. Sana *et al.,* "Performance analysis of machine learning techniques on software defect prediction using NASA datasets," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 300–308, 2019.

[33] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao *et al.,* "Developing a dengue forecast model using machine learning: A case study in China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 10, pp. 1–22, 2017.

[34] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam *et al.,* "Phishing web site detection using diverse machine learning algorithms," *Electronic Library*, vol. 38, no. 1, pp. 65–80, 2020.

[35] M. Alehegn, R. R. Joshi and P. Mulay, "Diabetes analysis and prediction using Random Forest, KNN, Naïve Bayes and J48: An ensemble approach," *International Journal of Scientific & Technology Research*, vol. 8, no. 9, pp. 1346–1354, 2019.