Gender Identification and Regression Analysis from Textual Facial Features: A Comparative Study

Faiaz Mohammad Tiham, Ehsan Abdullah Khan Saad, Nihad Adnan Shah Tirtho, Gazi Md. Julcarnine, Hossain Mohammed Usman, Sadiul Arefin Rafi, Md Humaion Kabir Mehedi and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{faiaz.mohammad.tiham, ehsan.abdullah.khan, nihad.adnan.shah, gazi.md.julcarnine, hossain.mohammed.usman, humaion.kabir.mehedi, sadiul.arefin.rafi}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—The paper explores gender classification and regression analysis using textual data of facial features. Various machine learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree, and Support Vector Machine (SVM), were implemented to predict gender and perform regression analysis. The classification models demonstrated strong accuracy, with KNN performing slightly better. On the regression side, KNN Regression and Linear Regression models showed lower Mean Squared Errors compared to SVM Regression. The paper's findings include the successful exploration of gender classification and regression analysis using facial feature data. By implementing various machine learning algorithms, accurate predictions of gender were achieved. These findings provide valuable insights into the application of machine learning for gender prediction and regression analysis.**

I. Introduction

Textual data has developed into a priceless resource for comprehending human behavior and preferences in the current digital era. People from many backgrounds are expressing their ideas, stories, and opinions on a wide range of subjects thanks to the growth of social media, online forums, and blogs. There are now more opportunities for research and analysis, particularly in the area of natural language processing (NLP), because of the explosion of online text information. In the NLP field, gender classification and regression analysis using textual data have become interesting research topics. These initiatives aim to use textual data's ability to reveal gender-related trends and insights. Understanding the gender of the people who create online material might have broad ramifications and uses, much like in the previous example of blog author gender classification. Weblogs, also referred to as blogs, are a distinct type of textual data. These online personal journals provide people with a relaxed setting for self-expression, experience sharing, and social interaction. In response to this trend, a plethora of research and commercial applications have mushroomed as the importance of blogs has expanded tremendously in recent years. The data gathered from blogs is invaluable for understanding public attitudes, preferences, and behaviors and may be used for everything from sentiment research to market intelligence.

In this regard, regression analysis and gender classification of blog authors show considerable potential. The gender distribution of blog authors can reveal information about the subjects that appeal to different genders most, the goods and services they like or dislike, and their general preferences. This information can be used for market-focused projects like product development and targeted advertising. Compared to more formal writings like essays or news stories, gender classification in blog postings poses special difficulties. Informality, conciseness, and the predominance of unstructured language define blog posts. They frequently have grammatical mistakes, slang, colloquialisms, and strange spellings. The work of gender classification is much more difficult due to these peculiarities. By utilizing a variety of characteristics and methodologies, previous research in this area has made an effort to address these difficulties. Even though progress has been made, this work proposes two unique strategies to improve gender classification and regression analysis of blog contributors even further. The first method uses pattern-based features to capture intricate aesthetic differences between male and female authors. These patterns provide a complex notion of authorship that is different from typical N-grams. The second method introduces a brand-new feature selection algorithm that makes use of a variety of selection standards and procedures. This strategy seeks to uncover the most discriminatory and illuminating traits for gender classification by integrating multiple selection approaches. Extensive studies are carried out utilizing real-world blog data obtained from a variety of blog hosting services to verify the efficacy of these strategies. The results highlight the potential of these advances in increasing gender categorization and regression analysis utilizing textual data by demonstrating considerable gains over current state-of-the-art techniques as well as publicly available systems.

In the sections that follow, we will go into greater detail about these methods, how they are used, and how they have affected the fields of gender classification and regression analysis, illuminating the wealth of information that can be gained by studying online textual content.

II. Related Works

A. Mukharjee et al. proposes two distinct ways to improve the gender categorization accuracy of blog contributors [1]. The first approach offers a new class of features by mining variable length POS sequence patterns from the training data using a sequence pattern mining method. The second technique is an innovative way for feature selection that combines several feature selection criteria and techniques. When compared to current state-of-the-art algorithms, the proposed strategies significantly improve classification accuracy, according to evaluation results using a real-world blog dataset. The report comes to the conclusion that additional research is required to examine how well the suggested strategies work on various categorization datasets. The proposed methodologies' increased accuracy compared to the current state-of-the-art systems is summated in one of the tables. According to the study, the suggested methodologies can be utilized to identify intricate stylistic patterns shared by male and female authors. The research also emphasizes how crucial it is to choose a subset of attributes that benefit from the classification task. To do this, the study suggests employing an ensemble feature selection method. Overall, the work offers a potential method for more accurately identifying blog authors' gender.

A. Bartle et al. introduces a brand-new method for employing deep learning models to categorize the gender of authors in diverse types of literature [2]. On datasets that included blogs, books, tweets, postings, and comments, the models were tested and trained. On the blog dataset, the WRCNN model had an accuracy of 86%. Additionally, the conventional NLP models that were employed in earlier attempts to categorize gender in authorship are discussed. Additionally, the file covers the model features, such as frequency measures, stylistic features, gender-preferential features, and POS pattern features. Additionally discussed are the dataset biases and the extraction procedure. With a split of 80% training, 10% development, and 10% test, the final dataset was divided into training, development, and test sets. The document offers a complete overview of the project and its results.

M. Martine et al. gives a summary of the 6th Author Profiling Task at PAN 2018, which concentrated on Twitter's multimodal gender identification [3]. The challenge was designed to investigate how author profiling may be accomplished using both text and picture data. Author profiling techniques have often depended heavily on textual data and conventional classifiers, necessitating considerable feature engineering. However, the PAN 2018 competition provided a chance to use picture content for the first time due to improvements in image recognition technology. The suggested method for the PAN 2018 challenge is built on cutting-edge deep natural language processing architectures in conjunction with a pretrained image classification architecture. Additionally, the paper makes mention of upcoming work on enhancing the image classification model as well as previous author profiling tasks.

K. Z. Haider et al. describes a real-time deep learning-based gender classification system for cell phones [4]. The tells about the difficulties and potential solutions for programming a computer to classify gender as correctly as humans. They suggest the Deepgender modified multilayer deep neural network, which combines 83 facial landmarks and a 3-D facial model for feature encoding and face alignment. The paper also explores similar work in the area of image classification as well as Deepgender's potential uses outside of gender categorization. The paper demonstrate a viable deep neural network design for handling very sizable facial picture datasets. The auto-tuning procedure and the datasets used for training and testing are also covered in the study. An open-source face database for researchers is the CAS-PEAL dataset. Deep gender, according to the paper, can be applied to a number of tasks other than determining gender, including racial binding, age determination, and expression identification. The study finds that Deepgender may be a helpful tool for smartphone picture processing and computerization.

F. T. Asr et al. introduce the Gender Gap Tracker, a computer program that analyzes text to detect gender bias in the media [5]. The method examines seven English-language Canadian news organizations' online daily publications to gain insights into the percentage of people named and quoted, broken down by gender, news organization, and study gender. The article analyzes the body of research on quoting patterns, information extraction from parsed language, and potential biases in identifying named entities' genders. The study contend that the first step in any attempt at change is a precise assessment of the current state of affairs. They contend that news organizations may be inspired by the Gender Gap Tracker to affect change in their sphere of influence. The processes taken to gather the data and process it in order to extract quotes, the speakers who made them, and their gender are summarized in the study. The Gender Gap Tracker can be used to address gender prejudice in the media and encourage more inclusive reporting, the paper say. They offer suggestions for possible uses of this technology in other spheres, including politics, healthcare, and education.

C. Bhagvati et al. explore the task of gender categorization for first names using deep learning [6]. It focuses on examining how word representations and deep learning architectures affect gender categorization performance. In the study, the effectiveness of well-known deep learning architectures, specifically LSTM and CNNs, in this task is compared. The studies show how gender classification performance is greatly impacted by the choice of word representation and deep learning architecture. The outcomes demonstrate that an LSTM design excels at this task. The study also examines the use of email and image analysis-based gender prediction systems, emphasizing earlier work in these fields.

S. Tilki et al. discuss Convolutions Neural Networks (CNN), in particular as a deep learning technique for object and character recognition in photos [7]. Deep learning automatically pulls pertinent features from the data, in contrast to conventional machine learning techniques that call for manual feature extraction. Deep learning has proven effective in a

number of industries, including classification, computer vision, optical character recognition, and object recognition. The CNN architecture's various levels, including the convolution layer, activation layer, pooling layer, fully connected layer, and dropout layer, are all covered in detail in the document. To teach the programme how to recognise objects, the training method entails feeding it photos of labeled objects. Since deep learning algorithms have a high level of classification accuracy, they are frequently applied.

H. Q. To et al. focus on gender prediction based on Vietnamese names using machine learning and deep learning techniques [8]. It begins with a literature review on previous studies and then describes the dataset and data collection methods. The paper presents a detailed description of the approach used for gender prediction and the experiments conducted. The results are analyzed using various machine learning models such as Support Vector Machine, Logistic Regression, and Long Short-term Memory. The paper concludes with a summary and suggestions for future work.

T. V. Janahiraman et al. discuss the impact of artificial intelligence (AI) on a variety of areas, including biometric systems, computer vision, and computer-human interaction [9]. It emphasizes the application of deep learning frameworks like TensorFlow for Asian face and gender classification. The popularity of deep learning architectures, in particular convolutional neural networks (CNN), for image classification applications is also mentioned in the document. It presents Keras as a high-level neural network application interface and TensorFlow as a commonly used machine learning framework. The article ends by highlighting the growing significance of deep learning in resolving complicated problems and comparing several models for gender classification.

B. Moghaddam et al. explore the problem of classifying gender from thumbnail faces without hair information [10]. It presents experiments comparing the performance of humans and machine classifiers, specifically Support Vector Machines (SVMs), in gender classification tasks using low and high-resolution images. The results show that human performance was adequate at high resolution but degraded with low-resolution images. SVMs demonstrated robustness and relative scale invariance for visual classification, outperforming human test subjects in both low and high-resolution tests. The paper also highlights the importance of hair cues in human gender discrimination and the impact of resolution on performance.

III. Dataset Description

Our research commenced with the acquisition of a comprehensive dataset retrieved from a CSV file, encapsulating vital information regarding facial features and corresponding gender classifications. An initial exploration of this dataset was carried out to gain a profound understanding of its inherent structure and characteristics. As part of this preliminary analysis, we opted to display the initial rows of the dataset to provide a snapshot of its contents, revealing a dataset comprising a substantial 5001 rows and 8 columns. In the course of this examination,We identified missing values in the 'forehead-width cm' (14) and 'forehead-height cm' (22) columns and addressed them through mean imputation. To ensure the integrity of our analysis and mitigate the impact of these data gaps, we employed a mean imputation technique. This approach allowed us to estimate and replace the missing values with the mean values of their respective columns, facilitating a more robust and complete dataset for our subsequent analyses. This meticulous data preprocessing step is crucial in ensuring the reliability and accuracy of our findings in the realm of gender prediction and regression analysis.

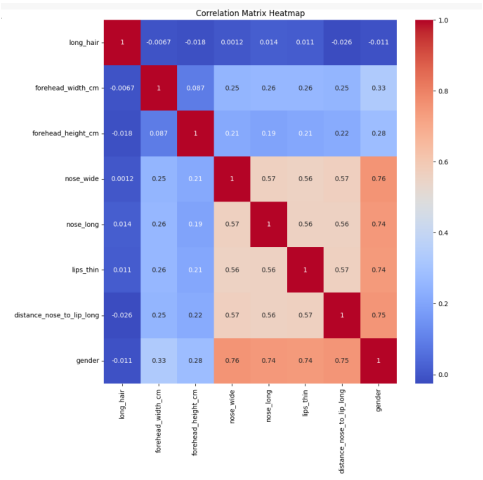

Fig. 1. Heatmap

IV. Methodology

A. Data Collection and Preprocessing

The study initiated with the collection of a dataset containing information pertaining to facial features and gender classifications, sourced from a CSV file. Upon dataset acquisition, a comprehensive data preprocessing phase was undertaken to ensure data integrity and suitability for machine learning analysis. Firstly, missing data was addressed through mean imputation, mitigating the impact of null values. To facilitate the integration of the 'gender' column into machine learning algorithms, a label encoding scheme was applied, where 'male' was encoded as 1 and 'female' as 0. Subsequently, the dataset was partitioned into feature variables (X) and the target variable (y) for both classification and regression tasks. Finally, a train-test split was executed, allocating data for model building and evaluation.

B. Exploratory Data Analysis (EDA)

Before embarking on model construction, an essential phase of exploratory data analysis (EDA) was conducted. To comprehend the distribution of the 'gender' column, a count plot was employed, revealing a relatively balanced distribution between male and female genders. In a bid to gain insight into the relationships among features, a pair plot matrix was constructed, offering visual representations of scatter plots depicting feature interactions.

C. Gender Classification Models

1) K-Nearest Neighbors (KNN): We initiated the classification task by implementing the K-Nearest Neighbors (KNN) algorithm. Through GridSearchCV, optimal parameters such as 'neighbors', 'weights', and 'algorithm' were determined. The best parameters were utilized to train the KNN model. Subsequently, predictions were made on the test set, and the model's accuracy was calculated. A classification report was generated, offering insight into precision, recall, and F1-score for each class. The confusion matrix was also visualized to comprehend the model's performance.

Decision Tree:

Next, a Decision Tree model was built. Using GridSearchCV, we identified the optimal criterion and 'max depth' for the tree. After training the model, we predicted gender on the test set and evaluated its accuracy. The classification report provided additional metrics, and the confusion matrix was visualized.

SVM:

Finally we have used SVM classification. SVM operates by transforming the input data into a higher-dimensional space (if necessary) and finding a hyperplane that best separates the classes. The hyperplane is chosen to have the largest margin between the nearest data points of the two classes. These nearest data points are known as support vectors. SVM can handle both linear and non-linear separation through the use of different kernels.

D. Regression Analysis Models

K-Nearest Neighbors (KNN) Regression:

The regression task commenced with K-Nearest Neighbors (KNN) regression. We trained a KNN regression model using the training data and predicted gender values for the test set. The Mean Squared Error (MSE) was calculated to quantify the model's prediction accuracy.

Linear Regression:

Linear regression was employed as a regression model. After training, predictions were made on the test data, and the MSE was computed to evaluate the model's performance.

Support Vector Machine (SVM) Regression:

SVM regression with a linear kernel was implemented. The model was trained, predictions were generated, and the MSE was calculated to gauge the accuracy of the SVM regression model.

E. Performance evaluation

Our performance evaluation underscores the proficiency of machine learning models, particularly KNN and Decision Tree for classification and KNN Regression for regression, in extracting meaningful patterns from facial feature data to predict gender and model gender-related attributes. These findings provide valuable insights into the application of machine learning in gender prediction and regression analysis based on facial features, with promising avenues for further research and real-world applications.

V. Model Comparison and Result Analysis

To gain a comprehensive overview of the classification models' performance, we generated a bar plot showcasing the accuracy of each model. This visualization aided in comparing the effectiveness of the KNN, Decision Tree, and SVM classifiers. We have got impressive results in all three classifiers. The accuracy of KNN classification is 97%. Again, the accuracy of the Decision Tree is 97%. Finally, the accuracy of SVM is 96%. So we can conclude that all of these classification models are good for dataset and KNN and Decision Tree are a little bit better than SVM classification.

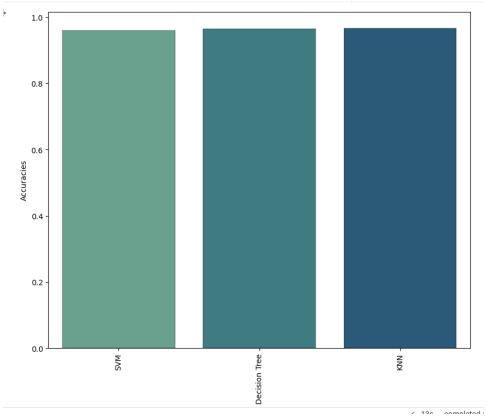

Fig. 2. Classification Comparison

A bar plot was generated to compare the Mean Squared Errors (MSE) of the different regression models. This visualization allowed for a clear understanding of how the KNN Regression, Linear Regression, and SVM Regression models performed in terms of prediction accuracy. We have got impressive results in all three regression models. The mean squared error of KNN regression is 0.029. Again, the mean squared error of linear regression is 0.0419. Finally, the mean squared error of SVM regression is 0.063. So we can conclude that KNN regression is better than the other two regression since it has the lowest mean squared error.

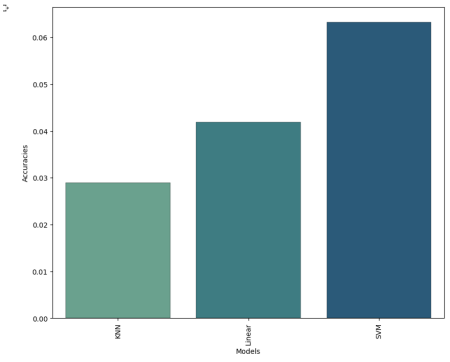

Fig. 3. Regression Comparison

VI. Conclusion

In conclusion, this project explored gender classification and regression analysis using facial feature data. By implementing various machine learning algorithms, we successfully predicted gender and performed regression analysis. The classification models (KNN, Decision Tree, and SVM) showcased strong accuracy, with the KNN model performing slightly better. On the regression side, the KNN Regression and Linear Regression models demonstrated lower Mean Squared Errors compared to SVM Regression. This project provides valuable insights into the application of machine learning for gender prediction and regression analysis.

References

[1] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, 2010, pp. 207–217.

[2] A. Bartle and J. Zheng, "Gender classification with deep learning," *Stanfordcs, 224d Course Project Report*, pp. 1–7, 2015.

[3] M. Martinc, B. Skrlj, and S. Pollak, "Multilingual gender classification with multi-view deep learning," in *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, 2018.

[4] K. Z. Haider, K. R. Malik, S. Khalid, T. Nawaz, and S. Jabbar, "Deepgender: real-time gender classification using deep learning for smartphones," *Journal of Real-Time Image Processing*, vol. 16, pp. 15–29, 2019.

[5] F. T. Asr, M. Mazraeh, A. Lopes, V. Gautam, J. Gonzales, P. Rao, and M. Taboada, "The gender gap tracker: Using natural language processing to measure gender bias in media," *PloS one*, vol. 16, no. 1, p. e0245533, 2021.

[6] C. Bhagvati *et al.*, "Word representations for gender classification using deep learning," *Procedia computer science*, vol. 132, pp. 614–622, 2018.

[7] S. Tilki, H. B. Dogru, and A. A. Hameed, "Gender classification using deep learning techniques," *Manchester journal of Artificial Intelligence and Applied sciences*, vol. 2, no. 2, 2021.

[8] H. Q. To, K. V. Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Gender prediction based on vietnamese names with machine learning techniques," in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 2020, pp. 55–60.

[9] T. V. Janahiraman and P. Subramaniam, "Gender classification based on asian faces using deep learning," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*. IEEE, 2019, pp. 84–89.

[10] B. Moghaddam and M.-H. Yang, "Gender classification with support vector machines," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 306–311.