Decoding Gender on Social Networks: An In-depth Analysis of Language in Online Discussions Using Natural Language Processing and Machine Learning

Gerasimos Vonitsanos*, Andreas Kanavos† and Phivos Mylonas†
*Computer Engineering and Informatics Department
University of Patras, Patras, Greece
mvonitsanos@ceid.upatras.gr
†Department of Informatics
Ionian University, Corfu, Greece
{akanavos,fmylonas}@ionio.gr

***Abstract*—In today's digital era, the internet is an indispensable platform for self-expression, facilitating communication, idea sharing, and community formation. Language, a pivotal tool in these online interactive spaces, is vital in reflecting personal identities, notably gender identification. This paper investigates gender identification on online discussion platforms, recognizing the crucial role of language in reflecting personal identities. The study employs Natural Language Processing techniques and machine learning algorithms to analyze data from a public discussion website. Beginning with a comprehensive literature review, the research explores the nexus between gender and language in online and offline contexts. The methodology involves data gathering, extensive preprocessing, and in-depth exploratory analysis, employing statistical methods and graphical representations. The study then rigorously evaluates their accuracy and effectiveness by applying diverse algorithms and models for gender-based text categorization. Results indicate the superior performance of transformer models, particularly distilBERT, in categorizing gender accurately. Additionally, the research underscores the challenges of gender-neutral analysis, emphasizing the need for inclusive methodologies in non-binary gender classification. The study contributes to the broader field of gender studies, providing valuable insights for future research and discussions on the interplay of gender and language in online spaces.**

***Index Terms*—Gender Identification, Natural Language Processing, Machine Learning, Text Categorization, Non-binary Gender Classification**

I. Introduction

Over the years, the rapid proliferation of Internet usage has transformed it into a ubiquitous platform for individuals to manifest their personalities, share life moments, and articulate their thoughts. Traditionally, online discussion forums were the primary arenas for such self-expression. However, the ever-evolving landscape of social media now offers users a multitude of avenues for self-representation. Consequently, this dynamic shift has ushered in an era of boundless textual information production [14], [15].

Concomitantly, computer science, particularly in Data Analysis and Machine Learning, has rapidly evolved, ushering in new technologies and approaches to language and text analysis. Natural language processing, a pivotal facet of Artificial Intelligence, centers on the capacity of computers to comprehend natural language and derive meaningful inferences. Beyond conventional Machine Learning algorithms, Neural Networks have emerged as an innovative and highly efficient approach for addressing challenges in natural language analysis from textual sources.

Given the vast reservoir of textual data available on the Internet and the cutting-edge tools of natural language processing, a compelling need arises to scrutinize users' language profiles and characteristics. This thesis embarks on the ambitious task of gender differentiation among users.

The research endeavors to explore the intricate relationship between gender and language within the context of online forums, with a primary emphasis on gender identification and expression. Through a meticulous analysis of linguistic patterns employed in these forums and applying text classification techniques, this study aims to provide deeper insights into how gender manifests and is portrayed through language.

Furthermore, this research aspires to construct a comprehensive and cohesive methodology for analyzing textual data in gender categorization, thereby contributing to the broader field of gender studies. It is poised to furnish valuable insights that will enrich future research endeavors and discussions about the interplay of gender and language within online spaces.

The rest of the paper is organized as follows. First, section II describes the relevant to the subject works. Besides, Section III analyzes the methodology followed, the algorithms, and the modules of our paper. Then, section IV details the implementation and different methods for practically evaluating our techniques. Moreover, in Section V, the acquired research results are captured. Finally, discussion and conclusions are outlined in Section VI.

II. Related Work

Language is a powerful tool that conveys information and reflects and shapes social norms, values, and identities. One

978-1-6654-8045-1/22/$31.00 © 2022 IEEE

area where language plays a central role is in the construction and expression of gender. The relationship between gender and language has been the subject of extensive research in sociolinguistics. Gender is a complex social construct that encompasses not only biological differences but also the cultural, social, and psychological dimensions that define what it means to be a male, female, or non-binary person in a given society. Language, as a means of communication and representation, plays a vital role in reinforcing, challenging, and shaping these gender norms and identities, and studying language and gender seeks to reveal the intricate connections between linguistic expression and the social construction of gender.

Gender and sex are increasingly being considered to improve scientific findings. Queer media studies study, on the other hand, underlines the pervasive misunderstanding and overlapping usage of the terms "sex," "gender," and "sexuality" among both laypeople and professionals [32]. The term "sex" primarily refers to the gender given at birth based on medical considerations (e.g., genitalia, chromosomes, and hormones), usually classified as 'male' or 'female,' and occasionally 'intersex.' Medical gender transition can change one's sex. The term "gender" refers to an individual's internal, firmly held sense of gender, also known as gender identity, which is impacted by social, cultural, and legal variables. The term "sexuality" refers to the physical, romantic, and/or emotional attraction to another person. We admit, however, that these definitions are socially created due to cultural expectations and conventions.

The interaction of these notions is complicated by modern civilization. Diverse socioeconomic groupings are acknowledged, recognized, legalized, and included in the complicated links between sex, gender, and sexuality by nation-states, governmental institutions, huge organizations, and corporations [27]. This intricacy extends to social media and digital platforms, where algorithms influence gender use, with automatic gender recognition systems as one example.

There is research that has extended the exploration of gendered language beyond traditional binary concepts. The study of Podesva and Roberts offers insights into the linguistic behaviors of trans people and reveals how language changes can accompany gender transitions, showing how language serves as a tool for expressing and reinforcing gender identity [26]. These findings underscore the importance of language in reflecting personal journeys of self-discovery and transformation.

In data mining, a substantial body of research has been dedicated to gender identification, mainly through analyzing textual data sourced from various social media applications. This burgeoning field represents a crucial intersection between data science and sociolinguistics, seeking to discern and understand how individuals express their gender identity through online communication. Researchers have explored diverse methodologies for accurate gender identification, ranging from linguistic pattern recognition to machine learning algorithms.

One notable approach involves the examination of linguistic features and writing styles employed by individuals across different social media platforms. Several studies have delved into the intricacies of language use, investigating how factors like vocabulary choice, sentence structure, and even emoji usage can contribute to the accurate identification of gender [22], [34]. Additionally, advances in natural language processing (NLP) techniques have significantly enhanced the precision of gender prediction models by enabling a more nuanced understanding of contextual language nuances [10], [20].

Scientists strive to offer users high-quality services as social networks and their members grow in number. Users' groupings can be used to show similar traits and interaction patterns amongst people participating in real-world activities. People who communicate with one another and are a part of connected communities make up social networks. One of the essential tasks of social network analysis is identifying these groups. Numerous methods for detecting communities have been developed to help locate intricate social network communities [11], [18]. A Twitter dataset was used for community detection in cultural and natural heritage data, reporting the analysis, with users rating community density. The research introduces a tailored methodology supporting multi-modal clustering and semantic annotations [16]. A novel methodology utilizing big data techniques was introduced to conduct extensive data analysis on Twitter, combining the Parallel Structural Clustering Algorithm for Networks (PSCAN) as a community detection algorithm and employing Latent Dirichlet Allocation (LDA) for topic modeling [8].

A widely favored approach for addressing the aforementioned challenge involves the application of Author Gender Detection (AGD) on texts within the realm of the Internet and cyberspace. AGD serves as a valuable tool for assessing the authenticity of individuals in social networks discerning instances of identity forgery. Recognizing the gender of individuals is crucial for analyzing their behavior on the Internet, providing insights into patterns and behaviors that contribute to a more comprehensive understanding of online interactions and dynamics [2], [31]. Using a dataset about the author's email recognition, the ANN was used to determine the gender of an email [7]. In another study, the gender of email authors was determined using a machine learning technique. A gender stratification classifier was created using a Support Vector Machine (SVM) by extracting linked email characteristics. The approach included features including morph-based, narrative, and emotive terms to improve the SVM's performance. The research findings proved the effectiveness of the suggested technique by showing that the SVM had an outstanding capacity to correctly identify the gender of the text's author [3].

III. Methodology

Machine Learning and Deep Learning are two fields of artificial intelligence that are now being applied in many scientific fields and everyday tasks. In this section, we briefly describe some of the most popular machine learning and deep machine learning algorithms, both from a general theoretical

point of view and an in-depth presentation of the mathematical foundations and approaches on which these models are based. The models and algorithms presented in this subsection were specifically chosen as they are also part of this paper's experimental process.

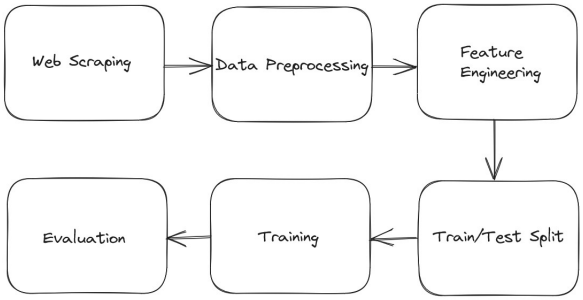

Fig. 1. Proposed Method

A. Logistic Regression

The logistic regression algorithm is a statistical learning algorithm mainly used in binary classification problems. It finds extensive applications in various fields, such as healthcare, finance, marketing, and social sciences. It is used for predicting medical outcomes, credit risk assessment, customer deviance prediction, sentiment analysis, and many other binary classification tasks. Its solid mathematical foundation and intuitive interpretation are powerful tools for understanding the relationships between input characteristics and target variables.

This algorithm is a generalized linear model that models the relationship between input attributes and the probability that the target variable belongs to a particular class. The fundamental principle of the algorithm is found in the logistic (or sigmoid) function, which maps the linear combination of input characteristics to a value between 0 and 1 [24].

The logistic regression function is defined as [13]:

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{1}$$

B. Naive Bayes

The Naive Bayes algorithm is a probabilistic machine learning algorithm commonly used for classification tasks. It is based on applying Bayes' theorem by assuming independence between features. Its applications are mainly related to text classification, for example, filtering spam messages and analyzing emotions through texts.

The core of the Naive Bayes algorithm is Bayes' theorem, which computes the posterior probability of a given class of observed features. Its mathematical expression is the following [25]:

$$P(C \vee X) = \frac{P(X \vee C) \cdot P(C)}{P(X)} \tag{2}$$

The Naive Bayes algorithm approach assumes that the features are conditionally independent given the class, which simplifies the calculation of the probability, which is represented as the product of the probabilities of individual features as [23]:

$$P(X \vee C) = P(x_1 \vee C) \cdot P(x_2 \vee C) \cdot \ldots \cdot P(x_n \vee C) \tag{3}$$

C. Random Forest

The Random Forest classification algorithm is a robust machine learning ensemble algorithm widely used for classification and regression tasks [1]. It combines the strengths of decision trees and ensemble learning to create robust and accurate models. Its ability to handle high-dimensional data nonlinear relationships and provide robust predictions makes it suitable for various real-world problems, including finance, healthcare, bioinformatics, and image recognition [6].

Regarding the mathematical approach of this model, Random Forest uses a set of decision trees to make predictions. Each decision tree in the forest is built on an initial sample of training data, and the final prediction is obtained by pooling the predictions of all trees [1].

Mathematically, the Random Forest prediction can be represented as:

$$Y = \text{mode}(Y_1, Y_2, \ldots, Y_n) \tag{4}$$

D. Support Vector Machines (SVM)

The SVM algorithm is also a machine learning algorithm for classification and regression, like the random tree algorithm. Its main areas of application are related to text classification, image recognition, and bioinformatics [5], and it is also used for sentiment analysis, object recognition, protein structure prediction, and credit scoring, among others. SVM's ability to handle high-dimensional data, its robustness against outliers, and its flexibility in capturing nonlinear relationships make it a popular choice in many real-world problems.

The Support Vector Machines model aims to find an optimal hyperplane that separates data points belonging to different classes by the maximum margin [4].

Mathematically, the SVM optimization problem is formulated as follows:

Minimize:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \tag{5}$$

subject to:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \tag{6}$$

and

$$\xi_i \geq 0, \quad i = 1, \ldots, n \tag{7}$$

where w is the weight vector, b is the bias term, C is the penalty parameter, x_i represents the input features, y_i is the class label, and ξ_i is the slack variable.

The optimization aims to find the decision boundary with the maximum margin while allowing for some misclassification (controlled by the parameter C) through the introduction of slack variables (ξ_i).

E. Neural Networks

Neural networks have revolutionised the field of artificial intelligence, enabling remarkable breakthroughs in areas such as computer vision, natural language processing, and speech recognition. They are computational models inspired by the structure and functionality of the human brain. They consist of interconnected layers of artificial neurons, also known as perceptrons, which process and disseminate information. Each neuron receives inputs, applies an activation function, and produces an output signal. The connections between neurons, represented by weights, are adjusted during training to enable learning and pattern recognition. The mathematical foundations of deep learning include linear algebra, calculus, and probability theory. Concepts such as matrix functions, gradient calculus, and optimization algorithms are critical in understanding and applying deep learning models. Activation functions, such as sigmoid, ReLU, and softmax functions, introduce nonlinearity and allow neural networks to approximate complex functions [12].

Deep machine learning is characterised by using neural networks with multiple layers. These networks can learn hierarchical data representations, extracting increasingly complex features at each level. In contrast, the depth of the network allows for more abstract and complex representations, thus capturing intricate patterns and the relationships between them [19]. Training in deep neural networks involves optimizing the network parameters to minimize loss while searching for the goal or training the model to return as valid results as possible. The widely used backpropagation algorithm calculates the slope of the loss function concerning the network weights and updates them accordingly. This process iteratively adjusts the weights to improve network performance. Stochastic Gradient Descent (SGD) and its variants, such as Adam's algorithms and RMSprop, are popular optimization algorithms used in deep learning [28].

F. Deep Learning

Deep learning has emerged as a revolutionary subset of machine learning, characterized by using neural networks with multiple layers to capture complex patterns in the data. The fundamental aspects of deep machine learning are discussed below, from neural network architectures to specific types such as Convolutional Neural Networks (CNN), Repeated Neural Networks (RNN), and transformer architecture.

At the heart of deep machine learning lies the neural network, inspired by the structure and function of the human brain. Neural networks consist of interconnected layers of artificial neurons, each process input data and passing information on to the subsequent layers. The layers typically include an input layer, one or more hidden layers, and an output layer. The weights and biases associated with the connections between neurons are learned during training, allowing the network to adapt to the data.

Deep learning leverages several building blocks to improve the performance and efficiency of the model. Activation functions, such as sigmoid, ReLU (Rectified Linear Unit), and tanh, introduce nonlinearity to neural networks, allowing them to capture complex relationships in the data. Backpropagation, a key concept, involves adjusting weights and biases by iteratively propagating errors backward through the network to minimize prediction errors [30].

Convolutional neural networks (CNNs) have revolutionized image analysis and computer vision tasks. CNNs are adapted for processing grid-like data such as images. They use convolutional layers to automatically learn and extract features such as edges, textures, and patterns from images. Convolutional layers reduce the spatial dimensions of data, preserving essential information while improving computational performance.

Recurrent neural networks (RNNs) excel in tasks involving series and time series data. RNNs retain memory of previous time steps, allowing them to record successive dependencies. However, traditional RNNs suffer from vanishing gradient problems, limiting their ability to capture long-range dependencies.

G. Transformers

In recent years, artificial intelligence and intense machine learning have seen the emergence of powerful models known as transformers, significantly enhancing natural language processing capabilities. Notably, two widely recognized transformers, RoBERTa Base and DistilBERT, have contributed substantially to various NLP tasks, including text classification, named entity recognition, and question answering.

1) RoBERTa Base Transformer: Developed by Facebook AI, RoBERTa Base is an optimized transformer model based on the BERT (Bidirectional Encoder Representations by Transformers) architecture. Renowned for its robust language understanding and representation capabilities, RoBERTa Base offers critical features that elevate its performance across various NLP tasks [21].

Key features of this transformer include:

- Contextual Word Embeddings: RoBERTa Base generates high-quality word embeddings by capturing the contextual information of each word within a sentence. This ability enables the model to comprehend intricate language nuances, encompassing word senses and syntactic structures, thereby significantly enhancing downstream task performance.
- Pre-trained Language Representation: Leveraging a substantial pre-training corpus, RoBERTa Base acquires comprehensive language representations. This initial pre-trained knowledge can be fine-tuned for specific downstream tasks, facilitating more effective and efficient training processes, especially with limited training data.
- Attention Mechanism: RoBERTa Base employs a self-attention mechanism, allowing it to focus on various segments of the input sequence during representation

building. This feature empowers the model to assign varying degrees of importance to different words, effectively facilitating critical contextual information capture.

2) DistilBERT Transformer: Hailing from Hugging Face, DistilBERT represents a compact version of the original BERT model, exhibiting comparable performance with significantly improved efficiency and reduced memory requirements. DistilBERT achieves this balance through a fusion of knowledge distillation and parameter reduction techniques [29].

Key features of DistilBERT include:

- Model Compression: DistilBERT employs knowledge extraction techniques to transfer insights from a larger, more intricate model, such as BERT, to a condensed version. This process enables the transformer to retain high performance while significantly reducing the model size and computational overhead.
- Faster Inference: Thanks to its streamlined architecture, DistilBERT delivers faster inference times than its larger transformer counterparts. This attribute renders it ideal for applications necessitating real-time or low-latency processing, enhancing its suitability for various real-world use cases.
- Efficiency and Resource Utilization: With reduced computational demands, DistilBERT offers a resource-efficient solution for researchers and practitioners with limited computing resources. It allows for efficient experimentation and development, even on devices with restricted computational power and memory capacity.

IV. Implementation

A. Twitter Discussion Synopsis

The dataset retrieved for this study was substantial, containing 6,627,794 tweets. The dataset was obtained using the Twitter API, which allows users to access a wealth of data from the platform, including tweets, user profiles, and more. Leveraging the Twitter API's capabilities, we were able to extract a diverse range of data from the platform, including user-specific information, textual data, and temporal information. Through targeted queries and filters, we gathered data from public discussions on a range of topics, enabling a comprehensive analysis of language patterns and gender identification in online discourse.

The dataset consists of the following features, each providing key information for the analysis:

- url
- username
- img
- gender
- text
- date

While the analysis and categorization primarily relied on the "gender" and "text" fields, the "username" field was also utilized to demonstrate some characteristic elements of the dataset.

Some important observations about the initial dataset are as follows:

- The number of tweets with the "gender" field blank: 143,840
- The number of tweets with an empty "text" field: 47,195
- The number of tweets with a blank "username" field: 0

After processing the data, the dataset was refined to 6,436,759 tweets. The "gender" data in the tweets comprises three distinct values: "male" for male gender, "female" for female gender, and "neutral" for individuals not identified by gender based on the data provided.

Regarding Regarding the gender distribution within the dataset:

- The number of male users is: 3,016,909
- The number of female users is: 2,743,347
- The number of gender-neutral users is: 676,503

Observing a well-balanced dataset in terms of male and female users, it's notable that the number of gender-neutral users is relatively smaller. To address computational limitations associated with training models on a large dataset, a subset of the data, consisting of 1,200,000 records where each class has 400,000 tweets, was employed for the experiment.

B. Data Analysis

Before proceeding with the text data cleaning process, an initial analysis of the word range, symbols, and noise distribution was conducted. The histogram in Figure **??** illustrates the distribution of word counts per text, emphasizing the prevalence of texts containing 0 to 100 words. Notably, the frequency sharply declines as the word count increases, with most entries remaining under the 500-word limit. Subsequent data preprocessing yielded a refined dataset, retaining clarity and significance for the experiment's analyses and outcomes.

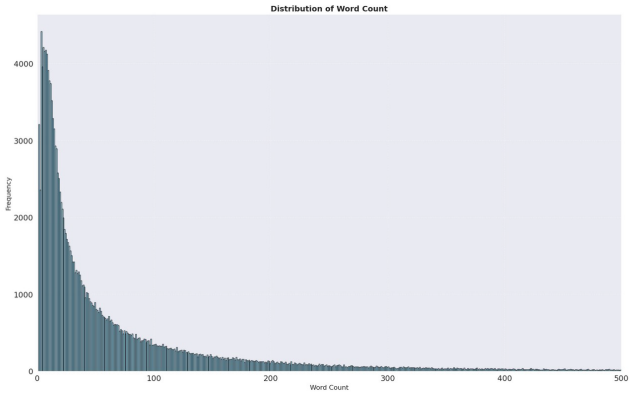

Fig. 2. Word Count Distribution

Upon close examination of the dataset, it became apparent that the extensive word count posed a considerable challenge during the text preprocessing phase. An in-depth analysis of the textual content and characteristics was imperative in ensuring the effectiveness of subsequent analyses. As part of the preprocessing step, the 500 most frequently occurring words were identified and subsequently removed, as depicted

in Figure 3. The elimination of these highly frequent, non-discriminative words such as "the," "I," "to," and "and" played a critical role in optimizing the training data, reducing the risk of overfitting and enhancing the models' ability to discern meaningful patterns and features [9].

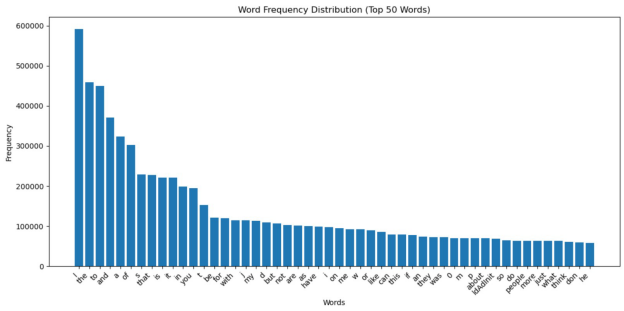

Fig. 3. Most Frequently Occurring Words in the Dataset

In the visualization provided in Figure 3, it is evident that certain common words dominate the dataset, with high repetition frequencies. The removal of these words significantly improved the dataset's quality, minimizing noise and enhancing the dataset's discriminatory features, thereby contributing to the accurate training of the models.

C. Data Preprocessing for Text Analysis

In text analysis and classification tasks, data preprocessing plays a pivotal role in ensuring the reliability and accuracy of the experimental results. Given the diverse and unfiltered nature of the dataset utilized in this study, it was imperative to conduct thorough data management procedures to eliminate noise and ensure data cleanliness. The subsequent steps outline the essential procedures in the data preprocessing phase, which helped streamline the dataset for practical classification model training and analysis.

The first crucial step in the data preprocessing pipeline involved the removal of numbers and symbols from the textual data. Numbers and symbols often introduce unnecessary noise and complexity, impeding the learning process and hindering model accuracy. By eliminating these elements, the textual information was simplified, reducing dimensionality and enabling the classification algorithms to focus on discerning meaningful patterns and relationships between words within the text.

Following the initial cleaning, the subsequent step entailed the elimination of web links, emails, and image links. Given the origin of the texts from online discussion forums, the presence of these elements was relatively high, contributing to the noise within the dataset. Removing these extraneous components not only enhanced the quality of the information but also improved the generalizability and interpretability of the classification models. This measure prevented overfitting and enabled the models to concentrate solely on text-relevant information, producing more accurate and robust outcomes.

Another critical aspect of the data preprocessing stage involved the removal of common words, known as "intermediate words" or stopwords, which typically lack significant contextual meaning, such as "the," "is," and "and." This step was crucial in reducing noise within the text, compelling the classification algorithms to focus on more meaningful patterns and relationships. Eliminating intermediate words also contributed to the interpretability of the classification models, ensuring that the models captured distinctive features to differentiate between different classes or categories in the text effectively.

The final step encompassed the application of lemmatization and stemming techniques to unify words and reduce variations. Lemmatization involved converting words into their primary forms while stemming aimed to reduce words to their root or base form by eliminating affixes. Both techniques contributed to the generalization of classification models, enabling the algorithms to comprehend the overall context of the text more accurately and make precise predictions and classifications.

V. Experimental Evaluation

After the comprehensive preprocessing and analysis of the dataset, the subsequent phase involved training a diverse set of algorithms and extracting results. This stage employed various machine learning models, encompassing classical algorithms, word embeddings, and transformer models, to perform binary and ternary gender categorization. The data inputs were categorized into four distinct text length ranges, enabling a detailed assessment of each algorithm's performance.

A. Text Length-based Analysis

The analysis of algorithm performance in relation to the length of the input text revealed intriguing insights. Results of the categorization accuracy for different text lengths are summarized in Table I. The experiments included the following text length categories:

1) Texts without word limit
2) Texts containing more than 20 and less than 100 words
3) Texts containing more than 40 and less than 80 words
4) Texts containing more than 30 and less than 150 words

Table I showcases the categorization accuracy for different text length ranges, indicating the performance of various algorithms. Notably, the transformer models, particularly distilBERT, demonstrate superior accuracy across all text length ranges, with an accuracy range of 64.88% to 72.43%. The fasttext algorithm also exhibits competitive performance, achieving an accuracy range of 62.31% to 65.02%. In contrast, classical algorithms, such as Logistic Regression, Random Forest, SVM, and Decision Trees, achieve relatively lower accuracy rates, ranging from 58.83% to 63.60%.

The classic algorithms encountered challenges in accurately predicting gender, with minimal fluctuations observed across different text lengths. In contrast, the fasttext algorithm exhibited improved results compared to traditional approaches. However, the transformer models, particularly distilBERT, demonstrated the most remarkable performance across various text length ranges. Notably, the compressed text lengths significantly enhanced the prediction accuracy of the algorithms,

TABLE I
CATEGORIZATION ACCURACY FOR DIFFERENT TEXT LENGTH RANGES

Algorithm	No word limit	20-100 words	40-80 words	30-150 words
Logistic Regression	59.44%	61.38%	63.60%	60.84%
Random Forest	59.16%	60.52%	61.32%	59.79%
SVM	58.83%	59.40%	59.82%	59.04%
Decision Trees	59.22%	60.03%	60.61%	59.80%
fasttext	62.31%	64.09%	65.02%	63.67%
RoBERTabase	63.60%	68.19%	71.12%	66.31%
distilBERT	64.88%	68.42%	72.43%	66.92%

highlighting the critical role of text length normalization in achieving reliable categorization results.

B. Gender-Neutral Analysis

Acknowledging the importance of gender-neutral categorization, the study addressed the complexities of identifying and classifying non-binary gender tags within the dataset. Given the relatively limited quantity of neutral gender data alongside the generalized nature of the dataset, all models, including the transformer models, faced challenges in accurately categorizing the neutral gender. This underscores the intricate nature of gender identification and the need for further research to address the nuances associated with non-binary gender classification.

C. Binary and Ternary Categorization with Neural Networks

The binary and ternary categorization experiments using the RoBERTabase and distilBERT models provided valuable insights into the performance of neural networks in gender classification. Despite the inherent challenges posed by the dataset, both models exhibited notable competence, demonstrating relatively superior performance compared to classical machine learning algorithms. The comprehensive results of the experiments are presented in Tables II, to IV.

TABLE II
CATEGORIZATION ACCURACY FOR DIFFERENT TEXT LENGTH RANGES (CLASSIC ALGORITHMS)

Algorithm	No word limit	20-100 words	40-80 words	30-150 words
RoBERTabase	62.12%	65.43%	68.47%	67.64%
distilBERT	63.77%	65.91%	69.09%	66.91%

TABLE III
CATEGORIZATION ACCURACY FOR DIFFERENT TEXT LENGTH RANGES (ADVANCED MODELS)

Algorithm	No word limit	20-100 words	40-80 words	30-150 words
RoBERTabase	64.19%	66.02%	68.12%	69.66%
distilBERT	65.71%	66.19%	68.93%	68.17%

Tables II and III present the categorization accuracy for different text length ranges, specifically focusing on the performance of the RoBERTabase and distilBERT models. These advanced models showcase robust performance, particularly in scenarios where the text length is constrained, with accuracy rates ranging from 62.12% to 69.66% for the RoBERTabase model and 63.77% to 68.93% for the distilBERT model.

TABLE IV
CATEGORIZATION ACCURACY FOR DIFFERENT TEXT LENGTH RANGES (BINARY CATEGORIZATION)

Algorithm	No word limit	20-100 words	40-80 words	30-150 words
Logistic Regression	51.12%	50.90%	51.32%	51.01%
Random Forest	50.06%	50.41%	50.73%	50.17%
RoBERTabase	55.27%	57.61%	58.45%	57.43%
distilBERT	54.98%	57.99%	59.09%	57.48%

Table IV delves into the categorization accuracy for binary classification, highlighting the performance of various machine learning algorithms. The results indicate that the RoBERTabase and distilBERT models exhibit relatively better accuracy rates, ranging from 54.98% to 59.09% and 55.27% to 59.09%, respectively, compared to classical algorithms like Logistic Regression and Random Forest, which demonstrate lower accuracy rates, ranging from 50.06% to 51.32%.

Overall, the results underscore the superior performance of transformer models, particularly distilBERT, in accurately categorizing gender in the dataset, highlighting the potential of deep learning models in addressing the complexities of gender identification. Additionally, the challenges faced during the gender-neutral analysis emphasize the need for further research and the development of inclusive methodologies for non-binary gender classification.

D. Importance of Gender-Neutral Analysis and Future Implications

The research emphasizes the critical importance of adopting a gender-neutral approach, emphasizing the need for inclusive methodologies that accommodate diverse gender identities. This study provides a foundational understanding of the complexities associated with gender identification in online discourse, thereby paving the way for further research and the development of robust frameworks for addressing the intricacies of gender representation in textual data.

The results highlighted the challenges and opportunities associated with gender identification in online discussion platforms, emphasizing the need for further exploration and refinement of methodologies for accurate and inclusive gender categorization.

VI. CONCLUSIONS AND FUTURE WORK

In conclusion, the paper discusses the evolution of online self-expression, particularly in the context of gender and language on social media. The study focuses on gender differentiation through text analysis, employing machine learning and deep learning techniques. It emphasizes the intricate relationship between language and gender, delving into the challenges of gender identification and expression in online forums. The research underscores the significance of gender-neutral analysis and the potential of transformer models, particularly distilBERT, in accurately categorizing gender. The findings call for inclusive methodologies to address the complexities

of gender representation in textual data, paving the way for future research and frameworks.

Future studies could explore multimodal approaches that leverage complementary information from different sources to enhance model accuracy [33]. Additionally, developing real-time gender recognition systems could have critical applications in various fields, including marketing, security, and social sciences. Investigating techniques to reduce inference latency while maintaining high accuracy will be critical to enabling real-time applications without sacrificing performance [17].

Expanding the range of gender identification models to include non-binary identities and identities with gender diversity is crucial in fostering the inclusion and recognition of all gender identities. Collecting and curating datasets that include a comprehensive representation of non-binary language and expressions and developing models sensitive to non-binary identifiers and pronouns will help encourage inclusive gender recognition systems that better reflect the complexity of human identity.

ACKNOWLEDGEMENT

This research was funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014-2020) under the Regional Operational Programme Ionian Islands 2014-2020, project title: "Indirect costs for project "Smart digital applications and tools for the effective promotion and enhancement of the Ionian Islands biodiversity"", project number: 5034557.

REFERENCES

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] B. Bsir and M. Zrigui. Bidirectional LSTM for author gender identification. In *10th International Conference on Computational Collective Intelligence (ICCCI)*, volume 11055 of *Lecture Notes in Computer Science*, pages 393–402. Springer, 2018.

[3] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[6] D. R. Cutler, T. C. E. Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[7] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu. Author gender prediction in an email stream using neural networks. *Journal of Intelligent Learning Systems and Applications*, 4(3):169–175, 2012.

[8] E. Dritsas, M. Trigka, G. Vonitsanos, A. Kanavos, and P. Mylonas. Aspect-based community detection of cultural heritage streaming data. In *12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–4. IEEE, 2021.

[9] E. Dritsas, G. Vonitsanos, I. E. Livieris, A. Kanavos, A. Ilias, C. Makris, and A. K. Tsakalidis. Pre-processing framework for twitter sentiment classification. In *Artificial Intelligence Applications and Innovations (AIAI)*, volume 560 of *IFIP Advances in Information and Communication Technology*, pages 138–149. Springer, 2019.

[10] A. Farzindar and D. Inkpen. *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2015.

[11] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[12] I. J. Goodfellow, Y. Bengio, and A. C. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, 2016.

[13] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009.

[14] E. Kafeza, A. Kanavos, C. Makris, G. Pispirigos, and P. Vikatos. T-PCCE: twitter personality based communicative communities extraction system for big data. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1625–1638, 2020.

[15] E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos. T-PICE: twitter personality based influential communities extraction system. In *IEEE International Congress on Big Data*, pages 212–219, 2014.

[16] A. Kanavos, M. Trigka, E. Dritsas, G. Vonitsanos, and P. Mylonas. Community detection algorithms for cultural and natural heritage data in social networks. In *Artificial Intelligence Applications and Innovations (AIAI)*, volume 628, pages 395–406. Springer, 2021.

[17] A. Kanavos, G. Vonitsanos, and P. Mylonas. Clustering high-dimensional social media datasets utilizing graph mining. In *IEEE International Conference on Big Data*, pages 3871–3880, 2022.

[18] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[19] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[20] J. Li, A. Sun, J. Han, and C. Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[22] L. Lopez-Santamaria, J. Gomez, D. L. Almanza-Ojeda, and M. A. Ibarra-Manzano. Age and gender identification in unbalanced social media. In *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 74–80. IEEE, 2019.

[23] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[24] S. Menard. *Applied Logistic Regression Analysis*. Number 106. Sage, 2002.

[25] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, 2012.

[26] R. J. Podesva, S. J. Roberts, and K. Campbell-Kibler. Sharing resources and indexing meanings in the production of gay styles. *Language and sexuality: Contesting meaning in theory and practice*, pages 175–189, 2002.

[27] V. Randall and G. Waylen. *Gender, Politics and the State*. Routledge, 2012.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[30] A. Savvopoulos, A. Kanavos, P. Mylonas, and S. Sioutas. LSTM accelerator for convolutional object identification. *Algorithms*, 11(10):157, 2018.

[31] A. G. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. B. Rybka, and T. Litvinova. Deep learning neural nets versus traditional machine learning in gender identification of authors of rusprofiling texts. In *8th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA)*, volume 123 of *Procedia Computer Science*, pages 424–431, 2017.

[32] C. Tannenbaum, R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger. Sex and gender analysis improves science and engineering. *Nature*, 575(7781):137–146, 2019.

[33] G. Vonitsanos, A. Kanavos, P. Mylonas, and S. Sioutas. A nosql database approach for modeling heterogeneous and semi-structured information. In *9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE Computer Society, 2018.

[34] K. M. Wagner, J. Gainous, and M. R. Holman. I am woman, hear me tweet! gender differences in twitter use among congressional candidates. *Journal of Women, Politics & Policy*, 38(4):430–455, 2017.