International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Advanced Document Representation Technique Based Approach for Author Profiles Prediction using Word Embedding Techniques

[1]D. Radha, [2]Dr. P. Chandra Shekhar

Submitted: 02/10/2023 **Revised**: 21/11/2023 **Accepted**: 02/12/2023

Abstract: The internet has become exponentially larger and unmanageably fast, the main contributing factor can be attributed due to how many people utilise social media, blogs, as well as online reviews. The most of the information given was published in various settings by multiple writers. The abundance of the information challenged Academics & information analysts collaborate to develop automatic methodologies for evaluating such content. Author Profiling is a widely employed approach in scholarly research, wherein scholars analyse the writing styles of authors to extract maximum information from texts. Author profiling is a methodology employed in the field of text categorization used to identify writers by their written works and forecast their demographic attributes, such as gender, age, native language, schooling, location, as well as personality traits. In today's information age, author profiling is a crucial approach with applications in forensic investigation, security, and marketing. Social media platforms have a substantial influence on our daily existence. and are a source of crimes, including public humiliation, fraudulent profiles, defamation, blackmail, stalking, etc. Author profiling helps the educational field by examining a big group of students. It aids in exposing the pupils' extraordinary potential. The educational forum also aids in determining the optimal amount of knowledge for individual students or groups of students. The majority of individuals of author profiling techniques employed a variety of Various criteria, encompassing linguistic factors, Various writing styles can be told apart by their content-based features, structure features, syntactic features, as well as semantic features. The present ones and models did there is no evidence to suggest that the enhancement of profile prediction accuracy has been achieved. They utilised new methods to improve the accuracy of demographic predictions for word embedding that are rooted in document analysis representation technique, it offers a new collection of style characteristics, feature selection algorithms, word weight measures, and the gender & age prediction models achieved accuracies of 0.9439 and 0.8945, respectively. The present study employs the PAN Competition 2014 evaluations database to do gender & age prediction. The experimental results are outperforming the earlier models and superior in estimation level.

Keywords*: Word Embedding Techniques, Term Weight Measures, deep larning, Gender Prediction, Age Prediction*

1. Introduction

The internet is an important factor to This paper aims to analyse the growth of social media, blogs, as well as reviews as observed in the literature [1]. The novel assortment of aesthetic attributes, algorithms for feature selection, measurements for term weight, & an innovative approach to document representations technique are most suitable for extracting information in different ways [2]. Currently, there is a growing demand for programmes that use machine learning, deep learning, as well as artificial intelligence to solve hard and important problems. [3]. The contemporary period is characterised by the prevalence of applications that have found utility in several domains such as marketing, security, and forensic investigation. [4]. Understanding the perpetrator's writing style using Author Profiling helps identify the offender [5]. The forensics examines writing styles, signatures, documents, and anonymous correspondence [6]. Customers were given a spot to evaluate the goods in the marketing field. The majority of the reviewers did not feel comfortable disclosing their identities [7]. In this particular case, those assessments were analysed in order to classify the customers according to variables including age, gender, job, language spoken at home, country, and personality traits [8]. Companies attempt to implement new business strategies to offer clients based on the categorization findings [9].

Knowing the author's background can be crucial. Forensic linguists, for example, would benefit from being able to determine the linguistic profile of an accused text message (the language most frequently employed by a specific group) in order to positively determine the sender [10]. Identifying criminals based on linguistic traits (language as evidence) alone from a textual analysis would be a huge assistance [11]. Furthermore, from a marketing perspective, businesses might be curious to learn, through the study of blogs & online product reviews, which demographics tend to favour or reject their items [12].

[1]*Department CSE, Gitam university, Visakhapatnam, India radharavavarapu@gmail.com*

[2]*Department of CSE , Gitam university, Visakhapatnam, India*

Chandrasekhar.pothala@gitam.edu

We've previously conducted a statistical analysis of Spanish-language usage across a variety of online mediums, including encyclopaedias, newsletters, blogs, forums, social networking sites like twitter and Facebook, and more [13]. In a recent study, we looked into whether or not linguistic cues were sufficient for detecting the six emotions. In pursuit of this objective, they created a collection of stylistic characteristics & attained commendable proficiency in discerning said emotions. We also analysed every potential permutation of gender, subject matter, and emotional tone in the phrase [14].

In this publication, we use the findings of our prior research to zero on the cognitive differences between the six and between people of different ages [15]. To this end, we propose a collection of traits for using stylistic characteristics to represent the work of unknown authors. We want to use this set of features to represent age and gender differences for usage in a machine learning setting [16]. A support vector machine approach employed in this study is trained & assessed utilising the PAN-AP-13 database. [17].

The findings are promising, but further examination of certain aspects is still required. The product itself is less important than the narratives you create around it in today's marketing landscape [18]. In this regard, the world is changing quickly, social media are expanding every day, & consumers are increasingly turning into active users in search of novel experiences. Therefore, it is becoming increasingly important to automatically scan the affective contents of social media in order to understand the desires & needs of the customers [19].

There is no denying the promotional, protective, and therapeutic value of social media. However, the reliability of the user-provided data might suffer if they do so. Numerous people invent their profiles' ages, sexes, organisations, and interests. One undeniable fact is that valuable insights may be derived from the content users generate and share on social media platforms. This presents a dual prospect for businesses and a formidable task for natural language processing technologies to comprehend users' demographic and psychosocial characteristics by analysing their writing style. [20].

Research conducted by Koppel, Argamon, and Shimoni (2003) has established a correlation between language utilisation and author demographics, specifically gender. However, it is important to note that the a lot of investigations in this area have mostly focused on the English language. A novel approach to the problem of analysing the emotional content of posts on social media is presented here in the form of an analytical method. Our working hypothesis is that a user's age & gender influence the manner in which they communicate their feelings regarding certain themes.

Utilising a methodology relying on graph theory, our goal is to model the manner in which people convey their ideas [21]. The capability of a graph-based technique to analyse complicated language structures is the primary driver behind its widespread adoption. The research conducted by Pennebaker (2011) serve as our primary source of inspiration. In Pennebaker's work, the style of writing is linked to human characteristics, such as demographics in the instance presented here. He used a collection of psycholinguistic the characteristics extracted from the texts encompass several linguistic elements, including parts of speech and phrases that convey emotional expressions, and so on [22].

When examining the writing style of individuals, it is important to consider how it integrate various parts of speech in a text, the verbs they use, the topics they talk about, as well as the feelings and attitudes they show (as well as their placement within the text), and other related factors [23]. According to Pennebaker's observations regarding gender and age differences, it is hypothesised that men tend to utilise a greater number of prepositions compared to women. This is attributed to their inclination to provide more detailed descriptions of their surroundings. Consequently, it is anticipated that men will employ a higher frequency of prepositional phrases in their writing, encompassing diverse subject matters and exhibiting varying levels of emotional expression. As a result, the sequence of preposition + determinant + noun + adjective is expected to assume a significant role in their linguistic compositions.

In accordance with this approach, they construct a graph that encompasses the many components of speech found within the user's texts. Furthermore, we enhance this graph by incorporating semantic information pertaining to the subjects they discuss, the manner in which they employ verbs, and how they make you feel [24]. The whole text is shown as a unified graph, encompassing punctuation marks to capture the writer's use of sentence structure and the connection of thoughts inside phrases [25]. After constructing the graph, several properties are derived from it and subsequently employed as features within a machine learning framework. [26].

The proliferation of offensive and discourteous discourse on online platforms represents merely one of numerous adverse consequences that have been precipitated by the swift ascent of social media in recent times. According to previous studies, communities of users that share the same assumptions are more likely to produce this type of offensive content [27]. Currently, the most advanced techniques for identifying instances of abuse solely rely on analysing textual data (i.e., lexical & semantic) clues

and thus are blind to user & community data [28]. In this work, they suggest a new way of thinking about this issue by utilising Twitter user profiles built by the community as a whole [29]. The experiments with a database of 16k tweets demonstrate that our algorithms greatly surpass how research as well as development are going in the area of abuse detection right now. Additionally, a qualitative investigation is conducted into the features of the model. All of the data & code utilized is made available to the public [30].

The goal of an author profile is to learn as much as possible regarding given author by a close reading of their published works. Age, gender, native language, education, and other socioeconomic indicators may all be used to narrow the focus of the investigation [31]. This is a profitable area with numerous uses in forensics, advertising, & network defence [32]. For instance, businesses may use textual analysis of product reviews to learn more about the kind of customers who like and dislike their items, while police departments may use such analysis to determine who committed a crime [33].

There has been a rise in interest in this field in recent years. Since then, an author-focused evaluation lab has been created as element of the PAN Workshop Series on finding theft, stealing other people's work, as well as misusing social software1 [34].

A set of blog texts was provided as input, and participants were tasked with devising methods to determine the gender (male or female) as well as age group (10s, 20s, or 30s) of the writers. Predicting classes is all about the profiles [35]. Over 70 teams registered for the task, and 21 of those teams actually participated by submitting an application. In PAN, participants are given texts whose authors' ages and genders have already been determined as training data, and their submitted programmes are then evaluated using a new, unseen dataset. Each entry is scored on how well it predicts the future [36]. The best scoring system correctly predicted gender 0.59 of the time and age 0.65 of the time. Author profiling is clearly a difficult undertaking with space for development [37].

Twitter is widely recognised as the most prominent tweeting service, where users may quickly & easily communicate and discuss information and ideas [38]. In the advertising industry, many of these Tweets are used to collect data and characteristics of Twitter users to zero in on subsets of extremely receptive audiences [39]. Sentiment analysis includes the subfield of "gender detection," which is concerned with determining a Twitter user's gender depends on how they write and other factors [40]. The purpose of this research is to examine the gender distribution of Twitter users by employing multiple Arabic classification mining techniques, including Naive Bayes (NB), Support vector machine (SVM), Naive Bayes Multinomial (NBM), J48 decision tree (DT), and K-Nearest Neighbours (KNN). When author names were included as a feature, the performance of the NBM, SVM, and J48 classifiers all increased to over 98%. In addition, the findings demonstrate that the pre-processing technique degrades gender recognition precision. The results of this study demonstrate that machine learning models are an effective means of identifying the authorship of a Tweet written in Arabic.

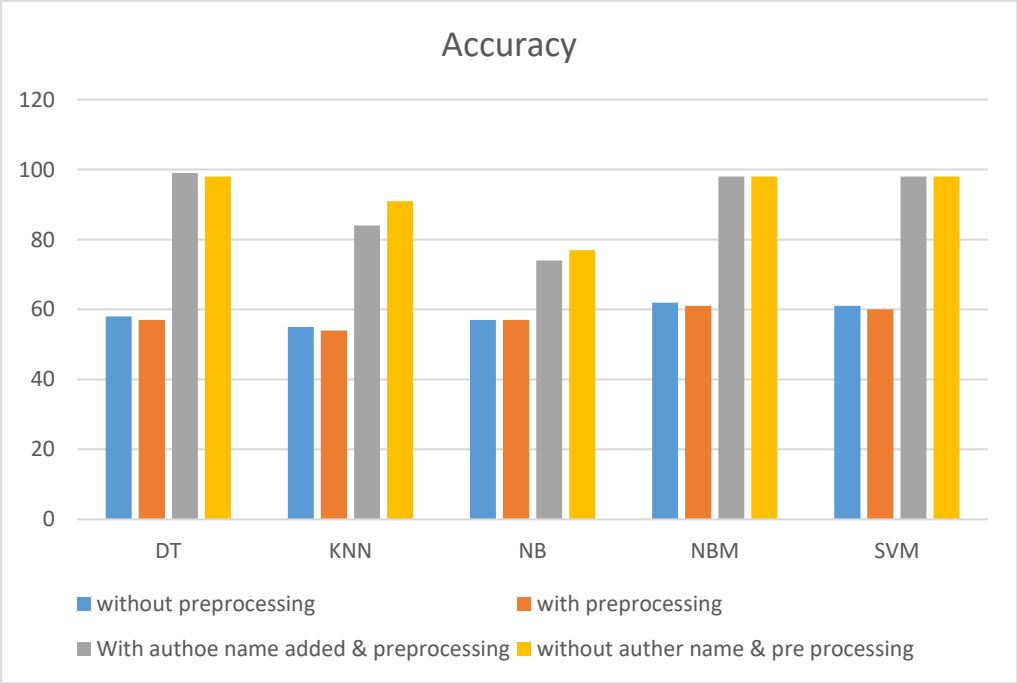


Fig 1: Performance of Classification models (With Author Name added).

When the author name characteristic was included in the database, the J48 classification achieved accuracy that was competitive with that of the NBM & SVM classifications. Our findings show that the J48 classifier's performance improves noticeably after the author name data is added to the database. As can be seen in Figure 1, the J48 & SVM classifications achieve a 98.69% accuracy with or without pre-processing, making them the most effective classifiers overall. This experiment yielded a total of 7929 successfully classified instances by J48 (with 3994 tweets from women and 3935 tweets from men), and SVM properly found 7929 instances (with 3990 tweets from women and 3939 tweets from men). We also find that both the J48 and SVM classifications did a good job of separating tweets from men as well as women.

2. Literature survey

Here, we conduct a quick literature review of recent studies. Koppel et al. [2003] is regarded as an early proponent of the author biography genre. They demonstrated that identifying trends in authors' writing styles may be used to categorise them into predetermined groups. In that research, the authors' genders and the texts' categories were determined by an examination of "meaningless" words (such as prepositions, pronouns, and auxiliary verbs) and how they impart meaning to the rest of the sentence through grammar.

After applying 1081 characteristics to More than 0.80, based on a sample size of 920 papers from the British National Corpus. Argamon et al. [2009] describe two fundamental classes of features that can be utilised in author profiling: content-dependent features (such as particular topics, keywords, and key phrases that are primarily used by one of the groups) and style-based features (such as average word length). Bayesian Multinomial Regression was used as the learning technique, and precision of 0.76 was achieved in experiments on texts from 19K blogs. Writers are more likely to use pronouns and determiners in female writing, while male writers are more likely to use prepositions and determiners. It was found that prepositions are more prevalent in more mature writing, while contractions without apostrophes were more characteristic of younger authors.

Mukherjee and Liu [2010] tackle the question of how to ascertain the sex of blog authors. They provide a feature selection method that makes use of POS tag pattern mining. A ten percent improvement can be achieved through feature selection, and a six percent improvement can be achieved through POS patterns, according to an experimental review.

Otterbacher et al [2010] likewise concentrated on gender classification, but using a different dataset this time: movie reviews.

This particular study focused on analysing the disparities in writing styles, subject matter, and information that exist between male and female authors (i.e., review count, review score, and reviewer rating). Such characteristics are incorporated into a classification based on logistic regression. When all of the different kinds of features utilised, that's when we got the best results. It is fascinating that they discovered that a metadata property, namely results were positive, the review was helpful, & male reviewers tended to be given higher marks than female reviewers.

Still on gender prediction, sarawgii et al. [2011] taken into consideration blogs & scholarly papers In Raghavan et al (2010) research, the author’s utilized probabilistic context-free grammars to identify syntactic regularities that went above shallow ngram-based features. Researchers observed that character-level models of language behaved significantly better than word approaches. As was to be expected, predicting gender in scientific articles was more difficult than doing so in blog posts.

The findings, however, demonstrate an accuracy of 61%, which is significantly higher than the probability of a random guess (which is 50%).This indicates that a person's gender may be determined even from their written work Predicting ages using many internet sources (blog posts, phone conversation transcripts, and forum posts) was the focus of Nguyen et al. (2011). In addition to the unigrams, POS tags, and word classes learned from LIWC, they also used gender as a defining feature. The use of linear regression in experiments has revealed that the combining of features produces the best outcomes. It was discovered that gender was an important factor in correctly identifying the ages of younger authors.

Peersman et al. [2011] Examined how well age & gender could be predicted on social networks. The brevity of the text is the most important factor in this case (12 words on average). Characters, word sequences (including unigrams, bigrams, and trigrams), & emoticons are some of the features that are utilised (bigrams, trigrams and tetra grams). A SVM classifier was used on those to determine their classification. It was discovered that emoticons and word characteristics were more helpful. The most successful outcomes were achieved when the training database was well-balanced.

Meina et al. (2013) used a wide range of structural aspects (the total number of phrases, sentences, & paragraphs), part-of-speech analysis, readability, emotion

words, emoticons, and statistical topics produced from (LSA) [Deerwester et al. 1990]. LSA uses a method called "dimension reduction," which is meant to reduce noise. There seem to be 311 features for age and 476 features for gender, which were used by a Random Forest classifier [Breiman 2001].

A Random Forest classification generates numerous decision trees & outputs the mode of their projected classes. The authors incorporated pre-processing, which had the purpose of finding spammy blog postings and removing them from the dataset. Take note of the fact that this was one of the slowest systems, requiring 4.44 days for analysing test data with the best accuracy (outside of training).

This approach is most accurate for gender & second for age.

In contrast to the majority of techniques, which describe documents as vectors by treating each term as a characteristic (also known as bag-of-words), suggest an alternative method for document representations called Second-order attribute. The crucial stage is computing how each phrase ties to each profile. The profiles indicate the classes whose results we want to predict: Female, male, 20s, 30s, and 10s. The term-profile relationship is primarily determined by frequency.

After the term vectors have been derived, the relationship among the document vectors as well as the profiles can then be determined. Their method accomplishes classifications into a total of six different categories. A LibLinear classifier (Fan et al. 2008) is utilised, and the attributes that it utilised were the top 50,000 most common phrases. Additionally, while classifying test data took 38 minutes, this method performed well. Age and gender accuracy were highest and third, respectively, with this method.

Table 1: Literature survey summary.

S No	**Author**	**Key points**	**Advanced model**	**Result**
1	Pennebaker et.al 2003 [7]	Men tend to use more determiners because they talk about real-world objects, whereas women tend to employ more first-person singular pronouns due self-conscious while speaking English.	LIWC,	Accuracy-75%-80%
2	Estival et al. 2008 [8]	Arabic emails	ACL	Accuracy 73.10%
3	Kholoud Alsmearat et al. 2015 [9]	Newsletters, Determine the gender of a book's author from the text itself; The issue is commonly referred to as "Gender Identity."	Supervised learning techniques	Accuracy 87.5%

4.	AlSukhni & Alequr et al 2016 [10]	Through the help of author names in Arabic tweets, a bag-of-words model was enhanced.	SVM,NB,NBM,KNN,	Accuracy-99.60%
5.	Burger et al 2011 [11]	employed statistical techniques to determine gender of users without revealing their identities; users came from all over the world and spoke many different languages.	WEKA to apply ML techniques like SVM, NB, balanced Winno2.	NB-68% SVM-72% balanced Winno2-75%

Mechti et al. [2013] calculated the top 200 most frequent phrases for each profile, then sorted them into several categories, such as types of determiners, types of prepositions, types of pronouns, types of words associated with love, types of terms frequently used by teenagers, and so on. There were 25 distinct varieties of English classes. The features were utilised to train a classifier based on a decision tree (J48). It took this system more than 11 days to categorize the test data, making it the slowest of the three. It came in at number two for predicting a person's gender, but it came in much further down the list for predicting a person's age.

We also took part in this competition by utilising ten characteristics that were dependent on data retrieval & two characteristics that were in accordance with readability assessments [Weren et al., 2013]. That was preliminary analysis, wherein we categorised only a limited amount of information gleaned from the instruction sessions.

3. Dataset Characteristics

Every year, experts in the fields of plagiarism detection, authorship attribution, authorship verification, & author profiling gather for the PAN (Plagiarism, Authorship, and Social Software Misuse) challenge. The competition's tasks evolve each year to throw a new set of variables at participants. The task of author profiling was first offered in 2013, encompassing several datasets and requiring the prediction of gender & age profiles. The database utilised in this study was sourced from collection of review scores from the 2014 PAN competition. The characteristics of the dataset are shown in Table 2.

Table 2: The PAN 2014 competition Reviews dataset

Classes / Profiles		The amount of reviews
Gender	Male	2090
	Female	2090
Age	18-24	370
	25-34	1000
	35-49	1000
	50-64	1000
	65+	900

The surge in the prevalence of author profiling is evident in the escalating participation rates observed across several author profiling competitions. Till 2019, the PAN organizers conducting competition on author profiling by changing the datasets and profiles need to predict. From 2019 onwards, they are conducting competition on A range of author profiling activities include celebrity

profiling, detecting of fake news spreaders, and identification of hate speech spreaders.

4. Proposed New Document Representation Methodology

Figure 2 shows the general steps followed in sugested model.

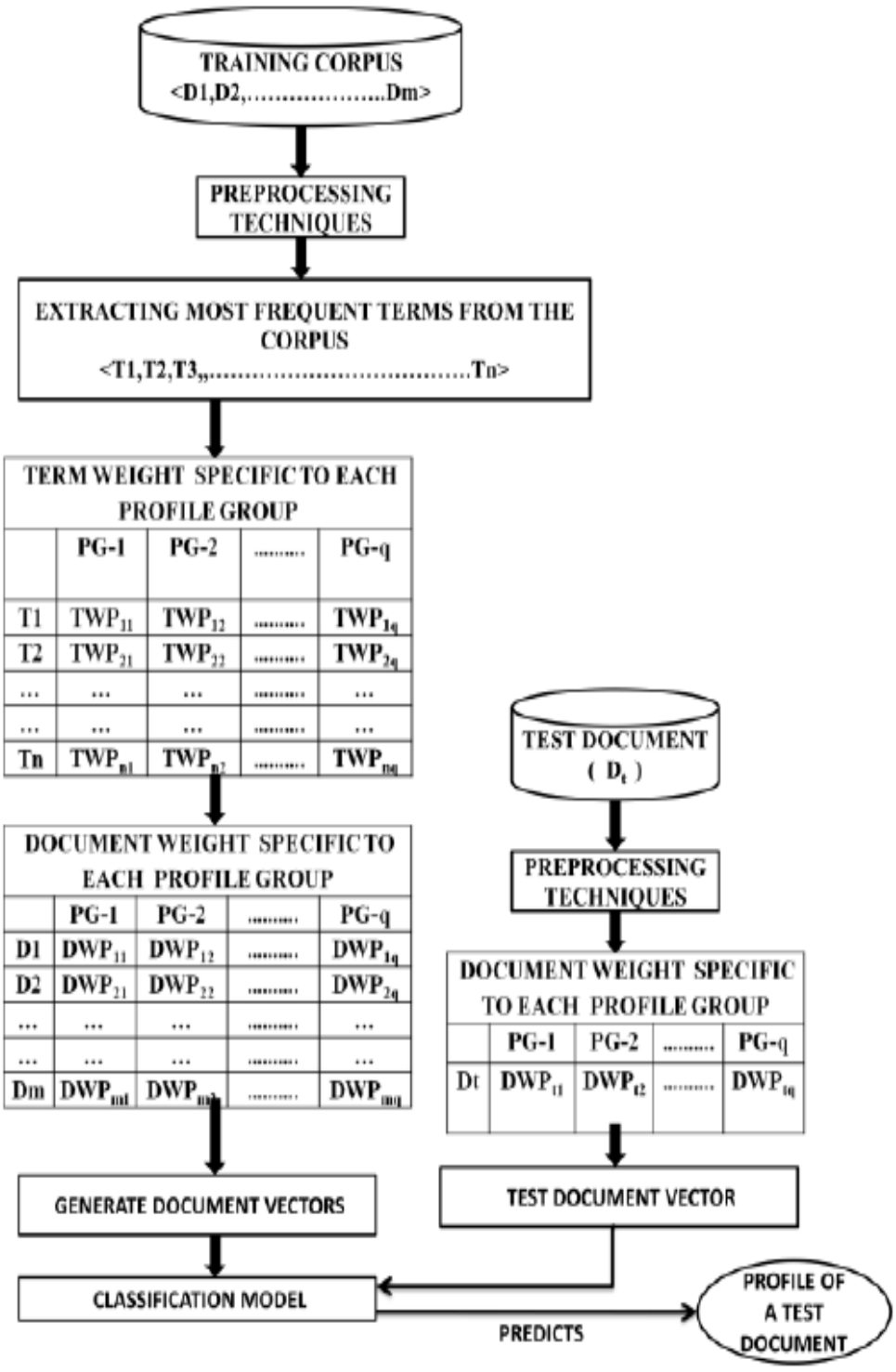


Fig 2: The architectural design of the suggested design.

The architectural design of the suggested model for Gender prediction is depicted in Figure 3.

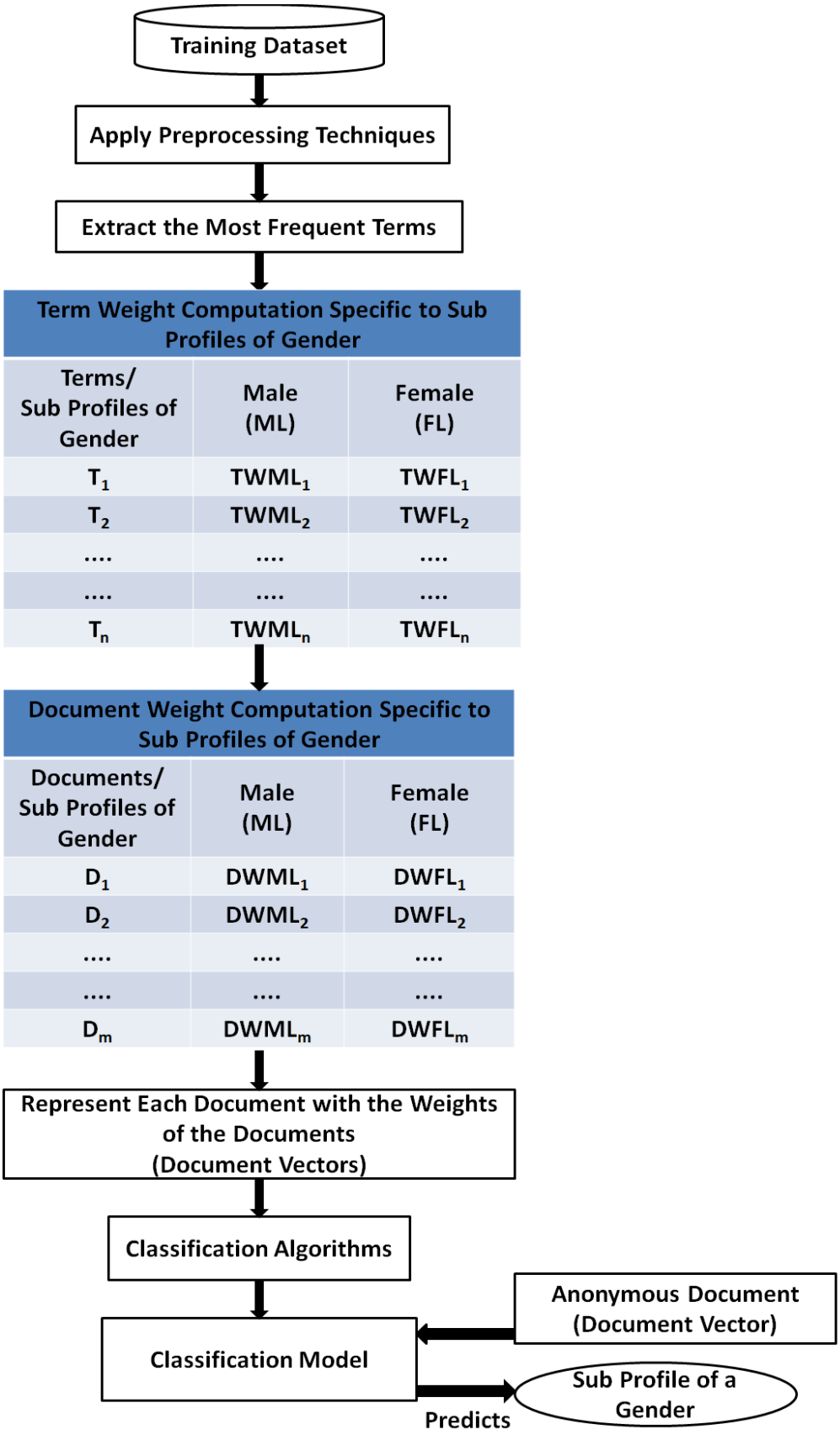


Fig 3: A Proposed Model for Predicting Gender.

The structure of the proposed model for predicting ageing is shown in Figure 4.

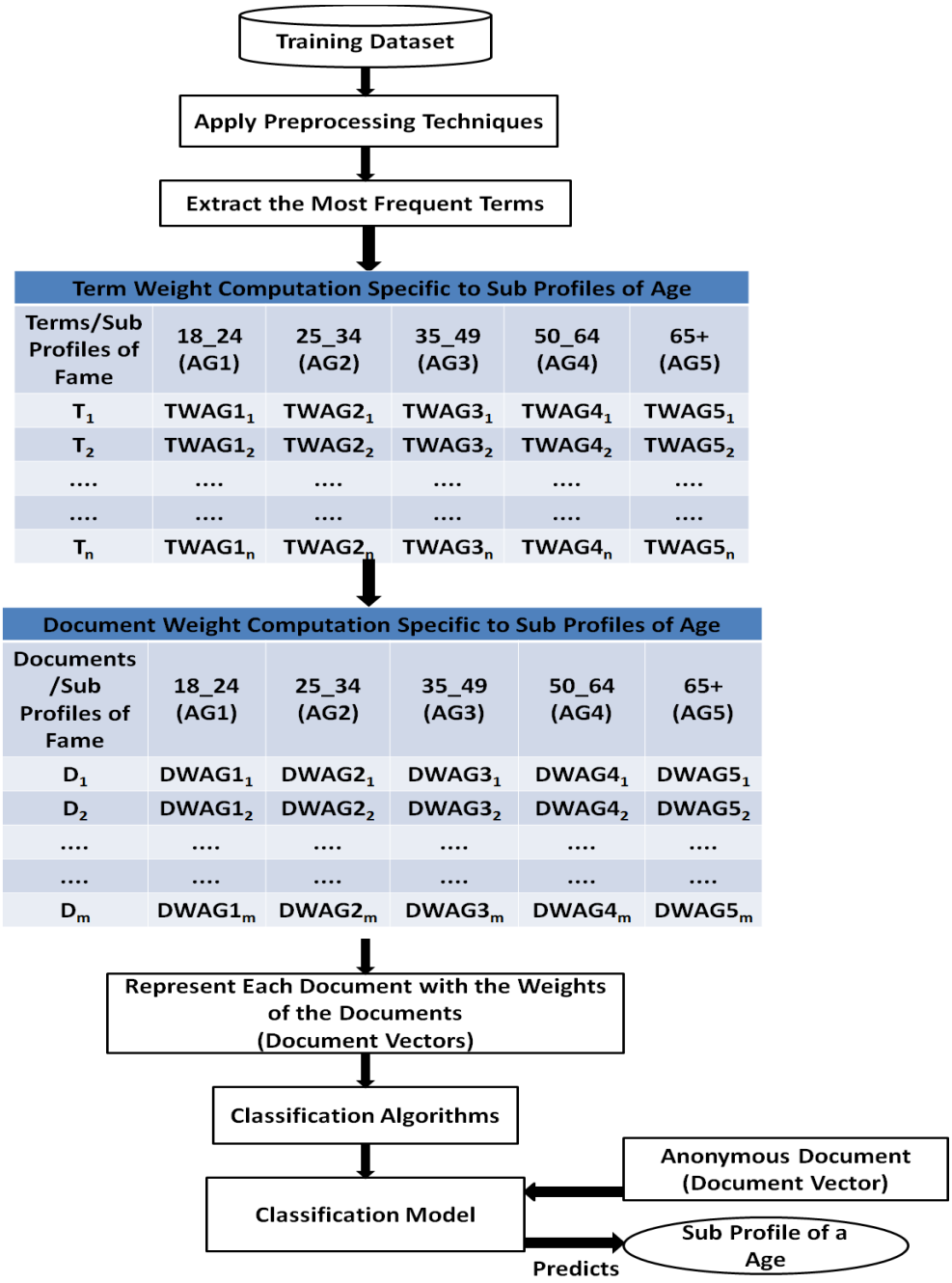


Fig 4: Proposed Model age prediction

4.1 Term Weight Measures

4.1.1 TFIDF (Term Frequency and Inverse Document Frequency)

$$TFIDF(Ti, Dk) = TF(Ti, Dk) \times log\left(\frac{N}{DF(Ti)}\right)$$

- $TF(T_i, D_k)$ is the frequency with which the phrase Ti appeared in document Dk,
- Let N represent the total number of documents inside the database,

4.1.2 The total number of documents in the total database that involve the term Ti is represented by the variable DF(Ti).

4.1.3 Term weighting is determined by evaluating the class density (CD) with respect to any and all papers falling under the same CDc.

$$CDc(Ti) = \frac{Nc(Ti)}{nc}$$

$$TW_{CDc}(T_i) = arg\max_{c}\left[CD_c(T_i)\right]$$

- $N_c(T_i)$ It is denoted by nc how many items in class c include the phrase Ti.

4.1.4 Term weighting utilising class density (CD) in relation to all articles inside the class (CDallc)

$$CDallc(Ti) = \frac{Nc(Ti)}{D(Ti)}$$

$$TW_{CDallc}(T_i) = arg \max_{c} \left[CD_{allc}(T_i) \right]$$

- The variable Nc(Ti) shows the total number of class c documents that contain Ti. Likewise, D(Ti) is the total number of documents in the entire dataset that include the phrase Ti.

4.1.5 TFRF (Term Frequency and Relevance Frequency)

$$TFRF(Ti, Dk) = TF(Ti, Dk) \times log\left(2 + \frac{A}{C}\right)$$

- Let A represent the number of positive class documents that contain the term Ti,
- C represent the number of negative class documents that contain the term Ti.

4.1.6 TF-Prob

$$TF - Pro\,b(Ti, Dk) = TF(Ti, Dk) \times log\left(1 + \frac{A}{B}\frac{A}{C}\right)$$

- A stand for the total number of class-positive texts that use the term Ti,
- B indicates the total number of class-positive papers that do not include the word Ti.
- C means how many papers in the "not positive" category contain the word "Ti."

4.1.7 TF-IGM (Term Frequency – Inverse Gravity Moment)

$$TF - IGM(Ti, Dk) = TF(Ti, Dk) \times \left(\frac{fi1}{1 + \lambda * \sum_{j=1}^{m} fij \times j}\right)$$

- The adjustable coefficient, denoted as λ, is varied within the range of 5 to 9. Typically, the default value assigned to the variable lambda (λ) is 7. The variable fi1 represents how many first-rate documents are indexed with Ti as a keyword. The number of classes, denoted by m, is a separate variable. Additionally, fij represents the fraction of papers belonging to the jth category that include the term Ti.

4.1.8 TF-IDF-ICSDF (Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency)

$$TF - IDF - ICSDF(Ti, Dk) = TF(Ti, Dk) \times \left(1 + log\left(\frac{N}{DF(Ti)}\right)\right) \times \left(1 + log\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{ncj(Ti)}{Ncj}\right)}\right)\right)$$

- Let N represent sum of all files stored in a given repository. The document's frequency (DF(Ti)) is the total number of documents in the collection that include the term Ti. The number of categories in the data collection is denoted by the m variable. To express how many documents in the jth class include the phrase Ti, we use the notation ncj(Ti). The total number of documents belonging to the jth class is denoted by Ncj.

4.2 Document Weight Measure specific to Profiles

$$Wdkj = \sum_{ti \in dk, dk \in pj} TFIDF(ti, dk) \cdot Wtij$$

$$TFIDF(ti, dk) = tf(ti, dk) * \boldsymbol{log}\ \left(\frac{|D|}{|1 + DFti|}\right)$$

- Wtij represents the importance of the word ti in the context of the profile group pj.
- The term frequency inverse document frequency (TFIDF) quantifies the importance of a given phrase ti within a given document dk.

4.3 Document Vector Representation

$$Z = \bigcup_{dk \in pj} (zk, cj)$$

$$zk = \{Wdk1, Wdk2, \ldots\ldots, Wdkq\}$$

- Wdki represents the mass of the dkth item in the pith profile.

5. A Methodology Utilising Word Embedding Techniques for Author Profiling

Figure 5 depicts the architectural framework employed in the Word Embedding Techniques depending Approaches for Gender Prediction.

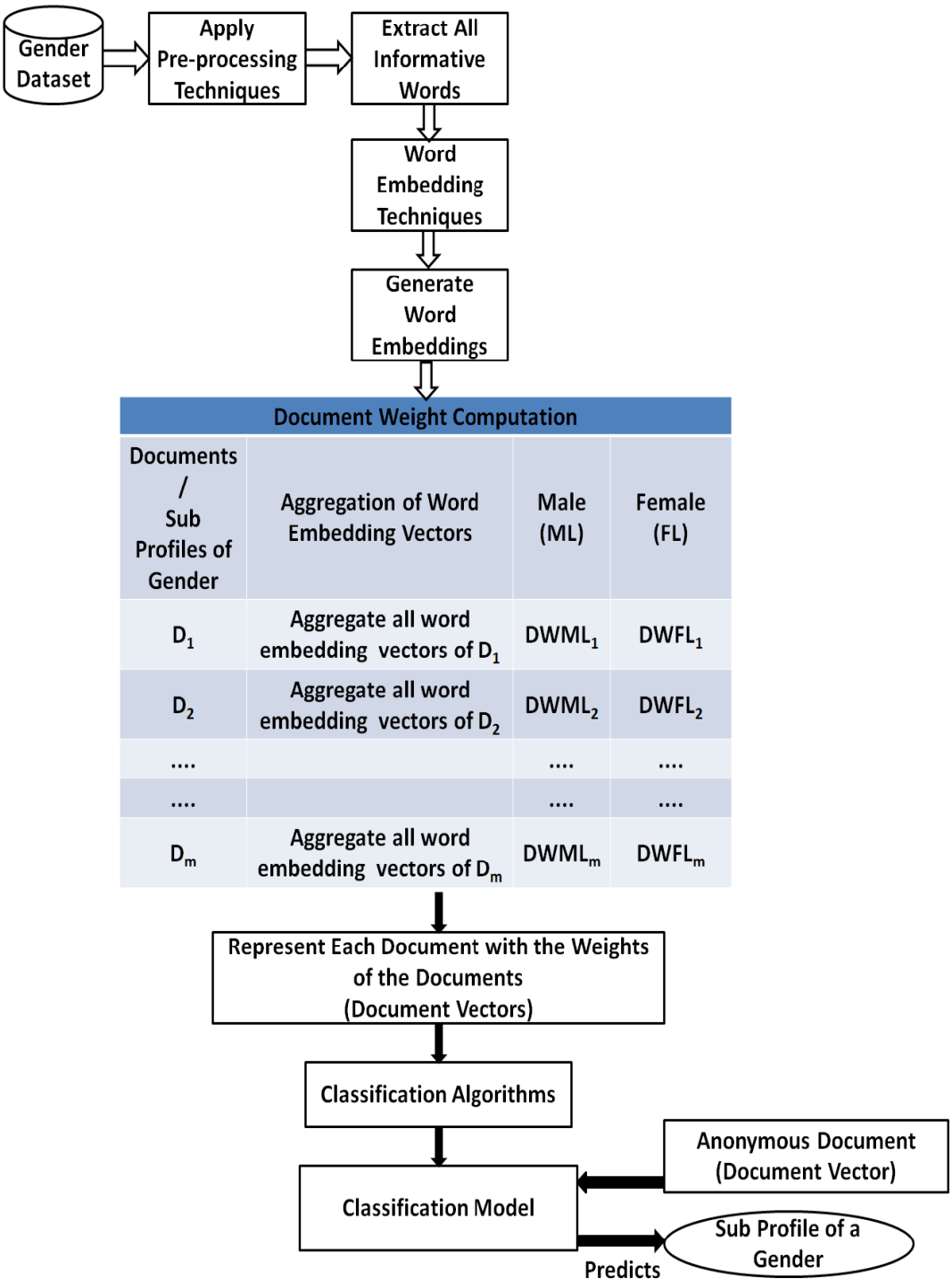
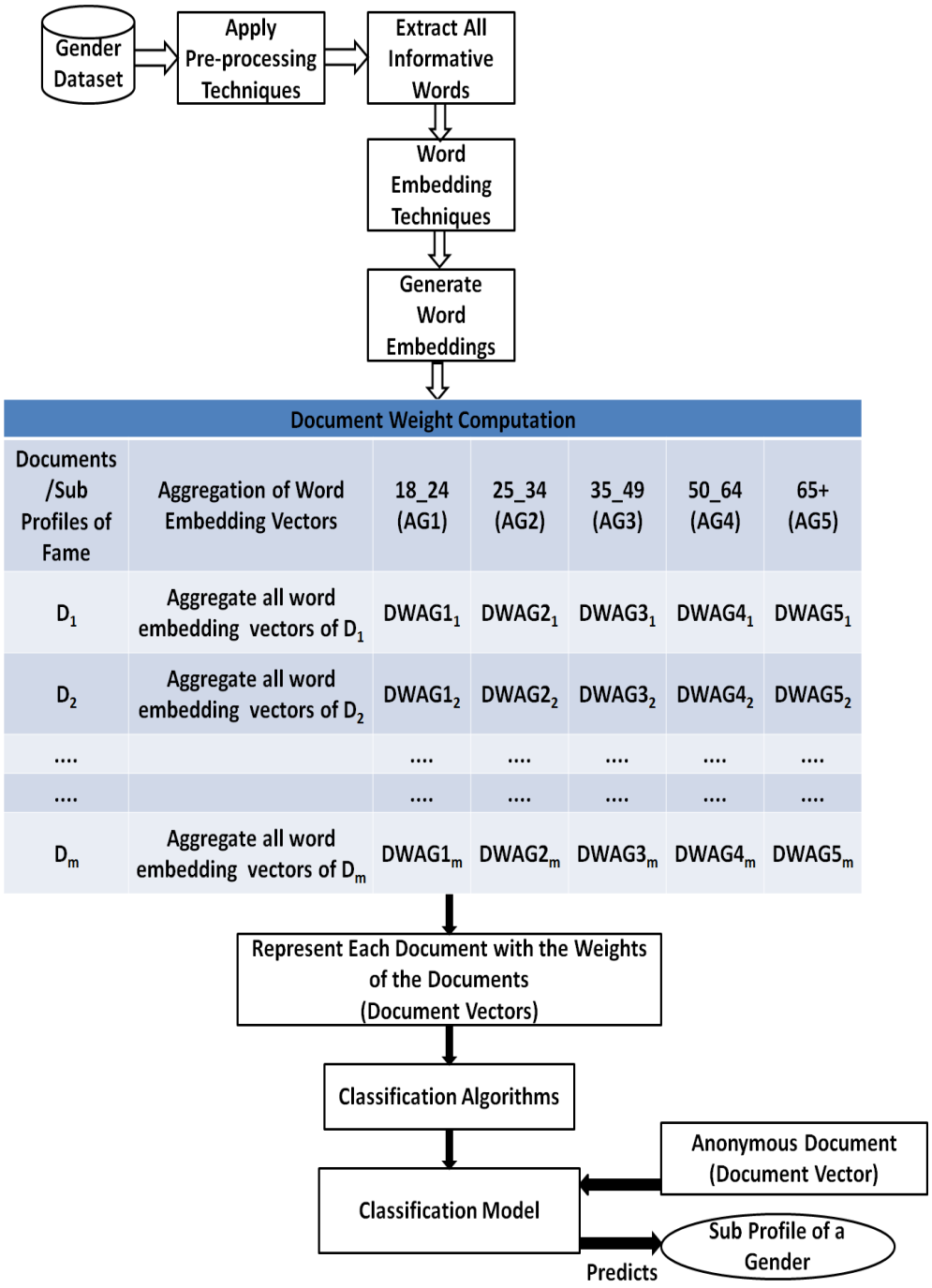


Fig 5 Illustrates a suggested method for gender prediction utilising word embedding methods.

The structure of the Age Prediction Approach utilising Word Embedding Approaches is depicted in Figure 6.



Document Weight Computation						
Documents /Sub Profiles of Fame	Aggregation of Word Embedding Vectors	18_24 (AG1)	25_34 (AG2)	35_49 (AG3)	50_64 (AG4)	65+ (AG5)
D_1	Aggregate all word embedding vectors of D_1	$DWAG1_1$	$DWAG2_1$	$DWAG3_1$	$DWAG4_1$	$DWAG5_1$
D_2	Aggregate all word embedding vectors of D_2	$DWAG1_2$	$DWAG2_2$	$DWAG3_2$	$DWAG4_2$	$DWAG5_2$
....		....	....	....	....	....
....		....	....	....	....	....
D_m	Aggregate all word embedding vectors of D_m	$DWAG1_m$	$DWAG2_m$	$DWAG3_m$	$DWAG4_m$	$DWAG5_m$

Fig 6: Method for Estimating a Person's Age via Word Embeddings.

Word Embeddings

- Word embeddings are typically regarded as word vectors with a defined length. These vectors are dense as well as dispersed, and they are formed based on the co-occurrence statistics of words.
- Word embeddings turn words into numerical vectors, known as "word vectors."
- The dimensionality of a vector is typically a major parameter of the chosen embedding technique.
- Word embeddings are categorised as “context-free” and “deep-contextual” varieties.
- Word co-occurrence and semantic links are used to build context-free word embeddings.
- In contrast to BERT's deep-contextual word embeddings, Word2Vec, fastText, and GLoVe produce "context-free" word embeddings.
- It is possible to use different word embedding strategies to produce many different numerical vector representations of the same text.
- Word embeddings are used to develop dense vectors of varying lengths and values from big textual collections.

- Word2Vec, for instance, has a vocabulary capacity of around 3 million words because it was trained on roughly 100 billion words from a sample of Google News.

6. Results and discussions

6.1 Predictions of age as well as gender for the new document illustration method

Table 3 shows the way the suggested new document format method predicts a person's gender when Support Vector Machine (SVM) classification is utilised.

Table 3: Performance of new document representation technique using SVM for Gender Prediction

NUMBER OF FEATURES	TFIDF	CDc	CDallc	TFRF	TF-Prob	TFIGM	TFIDF ICSDF
1000 TERMS	**68.32**	**73.90**	**79.40**	**79.25**	**80.40**	**83.25**	**83.48**
2000 TERMS	**69.54**	**75.29**	**82.20**	**82.15**	**81.20**	**84.15**	**85.17**
3000 TERMS	**71.10**	**76.48**	**84.35**	**84.60**	**81.35**	**85.60**	**86.43**
4000 TERMS	**71.31**	**77.31**	**85.55**	**86.35**	**83.55**	**85.35**	**87.20**
5000 TERMS	**71.78**	**78.17**	**87.00**	**87.80**	**84.00**	**87.80**	**87.73**
6000 TERMS	**73.15**	**78.87**	**88.40**	**89.75**	**86.40**	**88.75**	**89.43**
7000 TERMS	**73.29**	**80.39**	**89.65**	**90.70**	**86.65**	**90.70**	**90.43**
8000 TERMS	**75.38**	**81.16**	**90.45**	**91.50**	**88.45**	**90.50**	**92.37**

Table 4 shows the gender prediction accuracies of proposed new document representation technique when Random Forest (RF) classifier is used

Table 4: Performance of new document representation technique using RF classifier for Gender Prediction

NUMBER OF FEATURES	TFIDF	CDc	CDallc	TFRF	TF-Prob	TFIGM	TFIDF ICSDF
1000 TERMS	63.91	69.67	76.27	79.22	80.27	83.22	87.97
2000 TERMS	65.53	70.71	78.33	80.38	81.33	84.38	88.33
3000 TERMS	66.69	72.23	79.39	80.63	83.39	86.63	89.37
4000 TERMS	69.31	74.49	79.86	82.69	84.86	87.69	90.43
5000 TERMS	70.97	76.97	80.77	82.91	85.77	87.91	91.07
6000 TERMS	72.59	77.02	81.69	83.67	87.69	89.67	92.87
7000 TERMS	72.73	77.11	82.71	85.08	87.71	90.08	92.97
8000 TERMS	73.19	78.82	84.70	86.52	88.53	91.89	93.15

Table 5 shows the age prediction accuracies of proposed new document representation technique when Support Vector Machine (SVM) classifier is used.

Table 5: Performance of new document representation technique using SVM for Age Prediction

NUMBER OF FEATURES	TFIDF	CDc	CDallc	TFRF	TF-Prob	TFIGM	TFIDF ICSDF
1000 TERMS	61.19	65.21	68.56	69.78	70.95	73.71	76.43
2000 TERMS	63.38	65.37	70.65	71.36	72.58	74.37	78.67
3000 TERMS	64.42	67.91	71.67	72.36	73.66	75.33	79.83
4000 TERMS	64.85	68.04	73.31	74.41	75.59	75.42	80.45
5000 TERMS	65.51	68.39	74.30	75.29	76.33	77.28	80.89
6000 TERMS	66.07	70.47	78.65	76.67	78.61	79.64	82.69
7000 TERMS	66.59	71.91	80.68	80.40	79.65	80.40	82.82
8000 TERMS	68.26	73.49	81.73	82.08	80.72	82.18	83.31

Table 6 shows the age prediction accuracies of proposed new document representation technique when Random Forest (RF) classifier is used

Table 6: Performance of new document representation technique using RF classifier for age Prediction

NUMBER OF FEATURES	TFIDF	CDc	CDallc	TFRF	TF-Prob	TFIGM	TFIDF ICSDF
1000 TERMS	58.18	62.11	66.88	68.96	74.88	76.96	78.29
2000 TERMS	59.23	64.42	69.67	73.67	76.67	78.67	79.32
3000 TERMS	61.41	65.59	74.40	74.12	77.40	79.12	81.90
4000 TERMS	62.89	66.35	76.40	76.99	79.40	79.99	82.83
5000 TERMS	62.94	68.40	78.65	77.98	79.65	81.98	83.06
6000 TERMS	63.06	69.27	79.67	81.30	81.67	83.30	83.74
7000 TERMS	63.59	69.69	81.31	82.10	83.31	84.10	85.45
8000 TERMS	64.77	70.57	82.72	84.39	83.91	85.48	86.97

6.2 Accuracies of gender and age prediction for Word Embedding Techniques based Approach

Table 7 shows the gender prediction accuracies of different Word Embedding Techniques

Table 7: Performance of gender Prediction using word embedding techniques

Word Embedding Technique / ML Techniques	SVM	RF
Word2Vec	0.8468	0.8539
Glove	0.8557	0.8624
fastText	0.8643	0.8894
BERT	0.8721	0.8976

Table 7 shows that the RF classification with BERT embeddings was the most accurate at predicting gender, with a score of 0.8976.

Table 8 shows the gender prediction accuracies of proposed Word Embedding Techniques based Approach

Table 8: Results of a suggested method for predicting gender based on word embedding approaches

Word Embedding Technique / ML Techniques	SVM	RF
Word2Vec + Document Weights	0.8708	0.8818
Glove + Document Weights	0.8872	0.8947
fastText + Document Weights	0.9041	0.9226
BERT + Document Weights	0.9235	0.9439

In Table 8, the RF classifier with BERT embeddings and documents weights achieved a best accuracy of 0.9439 when guessing the gender.

Table 9 shows the age prediction accuracies of different Word Embedding Techniques

Table 9: Performance of age Prediction using word embedding techniques

Word Embedding Technique / ML Techniques	SVM	RF
Word2Vec	0.7984	0.8149
Glove	0.8059	0.8236
fastText	0.8147	0.8368
BERT	0.8239	0.8417

In Table 9, the RF classifier with BERT embeddings age forecast with a best accuracy of 0.8417.

Table 10 shows the age prediction accuracies of proposed Word Embedding Techniques based Approach

Table 10: Performance of age prediction method utilising Word Embedding Approaches.

Word Embedding Technique / ML Techniques	SVM	RF
Word2Vec + Document Weights	0.8267	0.8438
Glove + Document Weights	0.8479	0.8654
fastText + Document Weights	0.8534	0.8727
BERT + Document Weights	0.8716	0.8945

In Table 10, the RF classification with BERT embeddings and document weights was most accurate at predicting age, with an accuracy of 0.8945.

7. Conclusion

In this study project, an experiment was done to try to figure out things like the author's gender and age. This work suggests two ways using techniques like document representations for sex and age prediction as well as the word embedding technique. With document weights and word embedding methods, the best accuracy was 0.9439 for predicting gender as well as 0.8945 for predicting age. When compared to the performance of current author profile methods, the proposed method does the best job of predicting age and gender.

References

[1] Rangel, F., & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science, 177.

[2] Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. Information processing & management, 52(1), 73-92.

[3] Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2018). Author profiling for abuse detection. In Proceedings of the 27th international conference on computational linguistics (pp. 1088-1098). Association for Computational Linguistics (ACL).

[4] Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: Predicting age and gender from blogs. Notebook for PAN at CLEF, 2013(2).

[5] Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., & Wives, L. K. (2014). Examining multiple features for author profiling. Journal of information and data management, 5(3), 266-266.

[6] Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015, September). Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF (p. 2015). sn.

[7] Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology. (54): 547– 577.

[8] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007, September). Author profiling for English emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (Vol. 263, p. 272).

[9] Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., & Kanaan, G. (2015, November). Emotion analysis of arabic articles and its impact on identifying the author's gender. In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-6). IEEE.

[10] AlSukhni, E., & Alequr, Q. (2016). Investigating the use of machine learning algorithms in detecting gender of the Arabic tweet author. International Journal of Advanced Computer Science and Applications, 7(7).

[11] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G.,"Discriminating gender on Twitter". In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1301-1309, 2011.

[12] Reddy, D. H., & Sirisha, N. (2022). Multifactor Authentication Key Management System based Security Model Using Effective Handover Tunnel with IPV6. International Journal of Communication Networks and Information Security, 14(2), 273-284.

[13] Reddy, D. A., Shambharkar, S., Jyothsna, K., Kumar, V. M., Bhoyar, C. N., & Somkunwar, R. K. (2022, September). Automatic Vehicle Damage Detection Classification framework using Fast and Mask Deep learning. In 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.

[14] Kumar, A., Janakirani, M., Anand, M., Sharma, S., Vivekanand, C. V., & Chakravarti, A. (2022). Comparative Performance Study of Difference Differential Amplifier Using 7 nm and 14 nm FinFET Technologies and Carbon Nanotube FET. Journal of Nanomaterials, 2022.

[15] Vijetha, T., Mallick, P. S., Karthik, R., & Rajan, K. (2022). Effect of Scattering Angle in Electron Transport of AlGaN and InGaN. Advances in Materials Science and Engineering, 2022.

[16] Kumar, H., Prasad, R., Kumar, P., & Hailu, S. A. (2022). Friction and Wear Response of Friction Stir Processed Cu/ZrO 2 Surface Nano-Composite. Journal of Nanomaterials, 2022.

[17] Bhanuprakash, Lokasani, Soney Varghese, and Sitesh Kumar Singh. "Glass Fibre Reinforced Epoxy Composites Modified with Graphene Nanofillers: Electrical Characterization." Journal of Nanomaterials (2022).

[18] Jangam, N. R., Guthikinda, L., & Ramesh, G. P. (2022). Design and Analysis of New Ultra Low Power CMOS Based Flip-Flop Approaches. In Distributed Computing and Optimization Techniques: Select Proceedings of ICDCOT 2021 (pp. 295-302). Singapore: Springer Nature Singapore.

[19] Chandrasekaran, S., Satyanarayana Gupta, M., Jangid, S., Loganathan, K., Deepa, B., & Chaudhary, D. K. (2022). Unsteady radiative Maxwell fluid flow over an expanding sheet with sodium alginate water-based copper-graphene oxide hybrid nanomaterial: an application to solar aircraft. Advances in Materials Science and Engineering, 2022.

[20] Shareef, S. K., Sridevi, R., Raju, V. R., & Rao, K. S. (2022). An Intelligent Secure Monitoring Phase in Blockchain Framework for Large Transaction. IJEER, 10(3), 536-543.

[21] Koppula, N., Rao, K. S., Nabi, S. A., & Balaram, A. (2023). A novel optimized recurrent network-based automatic system for speech emotion identification. Wireless Personal Communications, 128(3), 2217-2243.

[22] Ali, F., Kumar, T. A., Loganathan, K., Reddy, C. S., Pasha, A. A., Rahman, M. M., & Al-Farhany, K. (2023). Irreversibility analysis of cross fluid past a stretchable vertical sheet with mixture of Carboxymethyl

cellulose water based hybrid nanofluid. Alexandria Engineering Journal, 64, 107-118.
[23] Pittala, C. S., Vijay, V., & Reddy, B. N. K. (2022). 1-Bit FinFET carry cells for low voltage high-speed digital signal processing applications. Silicon, 1-12.
[24] Saran, O. S., Reddy, A. P., Chaturya, L., & Kumar, M. P. (2022). 3D printing of composite materials: A short review. Materials Today: Proceedings.
[25] Dasari, K., Anjaneyulu, L., & Nadimikeri, J. (2022). Application of C-band sentinel-1A SAR data as proxies for detecting oil spills of Chennai, East Coast of India. Marine Pollution Bulletin, 174, 113182.
[26] Rao, A. D., Chaitanya, A. K., Seshaiah, T., & Bridjesh, P. (2022). An Integrated Approach by Using Various Approaches for a Green Supplier Selection Problem. In Recent Advances in Manufacturing, Automation, Design and Energy Technologies: Proceedings from ICoFT 2020 (pp. 909-919). Springer Singapore.
[27] Yakaiah, P., & Naveen, K. (2022). An Approach for Ultrasound Image Enhancement Using Deep Convolutional Neural Network. In Advanced Techniques for IoT Applications: Proceedings of EAIT 2020 (pp. 86-92). Springer Singapore.
[28] Kumar, C. A., & Haribabu, K. (2022). A Great Adaptive SNR Assumed Low Power LDPC Decoder. In Advanced Techniques for IoT Applications: Proceedings of EAIT 2020 (pp. 443-451). Springer Singapore.
[29] Thottempudi, P., Dasari, V. S. C. B., & Sista, V. S. P. (2022). Recognition of Moving Human Targets by Through the Wall Imaging RADAR Using RAMA and SIA Algorithms. In Advanced Techniques for IoT Applications: Proceedings of EAIT 2020 (pp. 544-563). Springer Singapore.
[30] Arun, V., Reddy, D. L., & Rao, K. N. (2022). A Novel Analysis of Efficient Energy Architecture in Cryptography. In Advanced Techniques for IoT Applications: Proceedings of EAIT 2020 (pp. 339-345). Springer Singapore.
[31] Amareswer, E., & Raju Naik, M. (2022). Smart Erobern of Vehicles on Crosswalks. In Advanced Techniques for IoT Applications: Proceedings of EAIT 2020 (pp. 489-497). Springer Singapore.
[32] Saikumar, K., Rajesh, V., Babu, B.S. (2022). Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal, Vol. 39, No. 1, pp. 31-42. https://doi.org/10.18280/ts.390104
[33] Kailasam, S., Achanta, S.D.M., Rama Koteswara Rao, P., Vatambeti, R., Kayam, S. (2022). An IoT-based agriculture maintenance using pervasive computing with machine learning technique. International Journal of Intelligent Computing and Cybernetics, 15(2), pp. 184–197.
[34] Saikumar, K., Rajesh, V. A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. Int J Syst Assur Eng Manag (2022). https://doi.org/10.1007/s13198-022-01681-7.
[35] Shravani, C., Krishna, G. R., Bollam, H. L., Vatambeti, R., & Saikumar, K. (2022, January). A Novel Approach for Implementing Conventional LBIST by High Execution Microprocessors. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 804-809). IEEE.
[36] Kiran, K. U., Srikanth, D., Nair, P. S., Ahammad, S. H., & Saikumar, K. (2022, March). Dimensionality Reduction Procedure for Bigdata in Machine Learning Techniques. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 836-840). IEEE.
[37] Srinivas Rao, K., Divakara Rao, D. V., Patel, I., Saikumar, K., & Vijendra Babu, D. (2023). Automatic Prediction and Identification of Smart Women Safety Wearable Device Using Dc-RFO-IoT. Journal of Information Technology Management, 15(Special Issue), 34-51.
[38] Sreelakshmi, D., Sarada, K., Sitharamulu, V., Vadlamudi, M. N., & Saikumar, K. (2023). An Advanced Lung Disease Diagnosis Using Transfer Learning Method for High-Resolution Computed Tomography (HRCT) Images: High-Resolution Computed Tomography. In Digital Twins and Healthcare: Trends, Techniques, and Challenges (pp. 119-130). IGI Global.
[39] Saikumar, K., Rajesh, V., & Rahman, M. Z. U. (2022). Pretrained DcAlexnet Cardiac Diseases Classification on Cognitive Multi-Lead Ultrasound Dataset. International Journal of Integrated Engineering, 14(7), 146-161.
[40] Maddileti, T., Sirisha, J., Srinivas, R., & Saikumar, K. (2022). Pseudo Trained YOLO R_CNN Model for Weapon Detection with a Real-Time Kaggle Dataset. International Journal of Integrated Engineering, 14(7), 131-145.