STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

Applying Author Profiling On Reddit Comments At The Document-Level

Idriss Oulahbib*, Meriem Benhaddi, Salah El hadaj

L2IS Laboratory, Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco

Abstract Author Profiling (AP) encompasses the task of discerning an author's biological, psychological, and socio-cultural attributes, including but not limited to gender, age, religion, profession, and personality, from their written content. This task is commonly approached as a form of text classification, where models are trained using features extracted from the author's text to predict labels such as gender and age category. This study investigates the effectiveness of Machine Learning (ML), Deep Learning (DL), and Transformer-based models for age and gender classification at the document level on a large dataset of Reddit comments annotated using Regular Expressions (REGEX). We employed various algorithms, including Naive Bayes (NB), Random Forest (RF), Logestic Regression (LR), Multi Layer Perceptrons (MLP), Convolutional Neural Networks 1 Dimension (CNN1D), and Distilled Bidirectional Encoder Representations from Transformers (DistilBERT). For feature extraction, we utilized Bag Of Words (BOW), Term-Frequency Inverse Document Frequency (TF-IDF), dictionary scores from Linguistic Inquiry Word Count (LIWC), averaged FastText embeddings (both pre-trained and trained on Reddit), and concatenated Subreddit embeddings to enhance contextual representation. Our experimental results revealed that traditional ML models with TF-IDF features, particularly LR, achieved competitive performance compared to deeper architectures. The best accuracy for gender classification was obtained by the DistilBERT + Subreddit embeddings model with 0.65 at the document level and 0.80 at the author level using majority voting. For age classification, the highest accuracy reached 0.37 with the same model configuration, outperforming all baseline approaches. These findings demonstrate that Transformer-based models enriched with contextual features offer a significant improvement over ML and traditional DL models in document-level AP.

Keywords Author Profiling, Document-Level, Machine Learning, Deep Learning, Reddit comments

DOI: 10.19139/soic-2310-5070-2762

1. Introduction

In the last decade, with the proliferation of online-generated content, especially on social media platforms, and advancements in Machine Learning (ML) and Deep Learning (DL) techniques, the field of Author Profiling (AP) has emerged and garnered significant interest among researchers in natural language processing. The large volume of online text has opened new avenues for analyzing not just the content itself, but also the authors behind the text, making AP a critical tool in deriving personal and socio-cultural characteristics from written content. Its applications span various domains, including psychology, social media studies, forensics [1], and marketing [2], where it has helped shape the understanding of human behavior and communication patterns.

AP involves identifying an author's personal and socio-cultural characteristics based on their written content. These characteristics often include demographic information such as age, gender, and profession, but the scope of analysis has broadened to include psychological traits and emotional states, adding complexity and depth to AP tasks. Researchers generally approach AP as a supervised learning task, where models are trained on labeled data to predict specific characteristics based on the linguistic features extracted from the author's text. As a result, AP has become a multidisciplinary field, intersecting with computational linguistics, sociology, and artificial intelligence, with significant implications for both academic research and practical applications.

*Correspondence to: Idriss Oulahbib (Email: i.oulahbib.ced@uca.ac.ma). L2IS Laboratory, Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1.1. Contribution

While most works have treated AP at the author level, which involves providing the AP model with multiple documents from the same author simultaneously, our work focuses on studying the applicability of AP for gender and age classification at the document level using a large Reddit comments dataset. We achieve this by employing ML and DL models, along with several feature extraction methods, including Bag Of Words (BOW), Term-Frequency Inverse Document Frequency (TF-IDF), the Linguistic Inquiry Word Count (LIWC) dictionary [18], Part-Of-Speech (POS) features extracted using NLTK, the average of pre-trained FastText word embeddings (original [19] and a version trained on our Reddit comments), embeddings from Distilled Bidirectional Encoder Representations from Transformers (DistilBERT), and subreddit name embeddings obtained from the original FastText model. This approach aims to provide a comprehensive evaluation framework that can serve as a valuable benchmark for short-text author profiling, addressing a gap in existing literature where the focus has primarily been on author-level classification.

1.2. Paper Structure

The rest of this paper is organized as follows: Section 2 provides an overview of the manual and automatic methods for AP data construction, along with a review of related work in age and gender AP. We explore various approaches employed in the literature, including traditional ML and modern DL techniques. In Section 3, we outline our methodology for constructing AP models at the document level, detailing the dataset, feature extraction techniques, and the ML and DL models implemented for age and gender classification. Section 4 presents the results obtained from our experiments, followed by a discussion of the model performances, including comparisons between different feature extraction methods and models. We analyze both the strengths and limitations of each approach, especially in terms of overfitting and generalization challenges. In Section 5, we conclude the paper by highlighting our key findings and discussing possible directions for future research.

2. Literature Review

2.1. Dataset Construction

The construction of a dataset for the AP task can be done through two methods:

- **Manual Annotation**: This method requires human annotators to manually extract the author's characteristics from their profile or written content. It is beneficial when working with smaller datasets, as human annotators can thoroughly analyze the nuances of the author's behavior and content, ensuring a higher quality of labels [10]. However, the manual nature of this process makes it less scalable for larger datasets.
- **Rule-Based Annotation**: This method relies on developing Regular Expressions (REGEX) to automatically extract author characteristics from text, making it more suitable for large datasets, especially those sourced from social media platforms. While this method is highly efficient for large-scale data, the quality of the labels may be impacted by certain factors, such as the presence of fake profiles or bots, which can distort the extracted information [7].

2.2. Related Works

Few studies have explored the influence of demographic characteristics such as gender and age on language usage. [3] examined the influence of age and gender on word usage in blogs on the Blogger platform. They found that females use more pronouns, assent/negation words, and blog-related vocabulary than males. In contrast, males use more articles, prepositions, and hyperlinks. Additionally, the study noted that teenagers tend to write about friends and mood swings, while older bloggers focus on topics like marriage, financial concerns, and politics. As bloggers age increase, their vocabulary shifts towards words related to money, jobs, and family, while references to sports, TV, and sleep decrease.

[4] statistically investigated the influence of gender on topics, language, and subreddit dominance on the Reddit platform. They found that Reddit is predominantly male-dominated, particularly in subreddits related to sports, video games, technology, and humor, while females dominate subreddits focused on beauty and specific TV shows. Women tend to discuss health, relationships, and personal experiences using positive language, expressing emotions, and writing longer comments. In contrast, men prefer topics such as sports, video games, technology, and humor, often using explicit language and writing shorter comments.

Several works have addressed the AP task, primarily focusing on constructing supervised learning models for classifying authors based on their demographic characteristics. In this section, we specifically discuss works related to gender and age classification tasks.

[5] conducted research on the b5-corpus [6], utilizing various feature extraction methods with a Logestic Regression (LR) classifier. The best performance they achieved ranged from 0.56 to 0.61 for age classification and 0.86 to 0.9 for gender classification tasks, using TF-IDF of the most frequent 3K terms.

[7] compiled a large dataset from Reddit publications, involving over 300K users across multiple categories, including age and gender. They trained several models at the author level, with their Hidden Attention Model (HAM) [8] achieving the best performances of 0.91 and 0.88, respectively, for gender and age classification using the AUC/ROC metric.

[9] focused on the EDGAD dataset [10], which comprises tweets from Egyptian Twitter users. They proposed a complex DL model named the Multichannel Convolutional Neural Networks (CNN) Bi-GRU model, achieving an accuracy of 91.37% for the gender classification task. Notably, they observed that the models performed better when fed with multiple tweets (e.g., 12 tweets) and had a larger maximum tweet length (e.g., 140 tokens per tweet).

[11] employed a heuristic algorithm to create a document representation by weighted averaging multiple feature types. Their approach was tested on PAN datasets for AP, spanning the years 2013 to 2018. The best performance was achieved on the PAN-15 dataset [12], with F1-scores of 0.90 for gender classification and 0.76 for age classification. Conversely, the lowest F1-scores were obtained on the PAN-14 dataset [13], scoring 0.68 and 0.18 for gender and age classification, respectively.

[14] applied several Large Language Model (LLM)s (including Polyglot, EEVE, and Bllossom) for AP in digital text forensics. Interestingly, the smaller Polyglot-1.3B model surpassed the larger EEVE-10.7B and Bllossom-8B models, attaining F1-scores of 0.84 for gender and 0.60 for age prediction.

The previously cited papers suffer from several limitations, including the construction of single-language models (e.g., English, Portuguese, Arabic) and a lack of generalization across various data sources (e.g., SMS, social media posts, instant messages). Recent works have begun to address these limitations [16] [15]. [16] focused on cross-genre AP by training a model on the Facebook corpus and testing it on Twitter and SMS corpora, and vice versa, as well as on code-switching AP, where documents contained both English and Roman-Urdu, for gender classification. Their model (Trans-Switch) achieved its highest accuracy of 74.07% when trained on the Facebook corpus and tested on Twitter, using a combination of RoBERTa and ULMFiT models.

3. Methodology

Our methodology, as illustrated in the flowchart (Figure 1), is divided into the steps detailed in the following subsections.

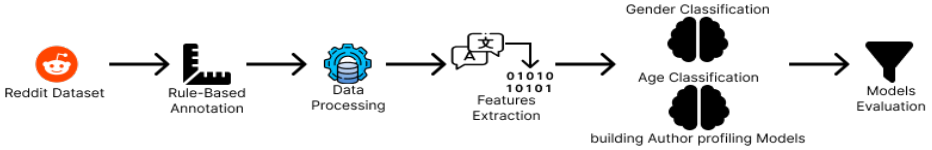


Figure 1. Methodology Flow-Chart

Table 1. Summary of Related Works

Ref	Dataset source	Data construction method	Features	Model	Performance measures	Limits
[5]	Facebook	Manual	TF-IDF	LR	F1-score: Age: 0.56 – 0.61, Gender: 0.86 – 0.9, Religiosity: 0.17 – 0.67, IT Background: 0.64 – 0.8	- Trained the model on an imbalanced dataset - Trained only on Facebook posts
[7]	Reddit	REGEX based	Word2vec	HAM	AUC/ROC: Age: 0.88, Family Status: 0.9, Gender: 0.91, Hobby: 0.80, Profession: 0.85	- Single language model - Trained only on Reddit data
[9]	Twitter	Manual	Word Embedding	Multichannel CNN Bi-GRU	Accuracy: Gender: 91.37%	- Trained only on the Egyptian dialect - Trained only on Twitter data
[11]	PAN datasets between 2013 and 2018	NA	Document Embedding created based on a genetic approach	Random Forest (RF), SVM, Extra Trees, KNN	F1-score: Age: 0.76, Gender: 0.9	- Single language model - Could be computationally heavy since it aggregate multiple features extraction methods
[14]	NIKL Korean Dialogue Corpus 2022	Manual	Word Embedding	Polyglot EEVE Bllossom	F1-score: Age: 0.60, Gender: 0.84	- limits on distinguishing adjacent age groups - Polyglot-1.3B model showed bias
[16]	Facebook, Twitter, SMS	Manual	Word Embedding	RoBERTa-ULMFiT	Accuracy: Gender: 74.07%	- Gender AP only - The generalization performance need to be improved

3.1. *Dataset & Annotation*

We utilized a Reddit comments dataset containing approximately 1.7 billion comments published between 2007 and 2015, with an uncompressed size exceeding 800 GB. To annotate the dataset, we adopted the approach[†] outlined in [17], which focuses on identifying specific demographic characteristics such as age and gender. This was achieved using REGEX, as detailed in Table 2, to systematically extract the relevant information from the comments.

Table 2. REGEX used for the data annotation task [17]

Demography Characteristic	**REGEX**
Age	.*?(i am—i\’m) (\\d+) (years—yrs—yr) old[ˆe].*?
Gender	Statements in which individuals refer to themselves using terms such as ’boy,’ ’man,’ ’male,’ or ’guy’ for males, and ’girl,’ ’woman,’ ’female,’ or ’gal’ for females (e.g., ’I’m male’ or ’I am a girl’).

For identifying the age, the REGEX search for sentences like ”i’m <number> years old”. To reduice the bias related to the fact that the users might lies regarding their age or gender. we used used the same heuristic annotation filtering as in [17].

3.2. *Data Processing & Features Extraction*

The objective in this phase is the cleaning of the comments and conversion from textual format to numerical format using common feature extraction methods for textual data.

- **Data size**: We used only a subset of the dataset containing the comments published within the 2007 to 2010 time span for model construction and evaluation (see Table 3).
- **Cleaning processes**: Removal of hyperlinks, and removal of empty comments. We did not apply any further cleaning, such as removing stopwords, punctuation, lemmatization, etc. We used the NLTK word tokenizer[‡] for the comments tokenization. Since the dataset is imbalanced, we employed a downsampling strategy to balance the dataset.

[†] https://github.com/cfwelch/compositional_demographic_embeddings
[‡] https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

- **Feature extraction methods**: We employed TF-IDF, BOW, the LIWC dictionary [18], and POS features extracted using NLTK. Additionally, we used the average of pre-trained FastText word embeddings (both the original version [19] and a version trained on our Reddit comments), as well as embeddings derived from DistilBERT. For subreddit name embeddings, we used the original FastText model.
- **Max features**: For controlling the number of features extracted by the TF-IDF and the BOW methods, we took the top 100, 250, 500, 1000, 5000, 12000, 20000, and 50000 features chose using Chi2 (χ^2) features selection method for selecting the most discriminative features for each task.
- **Train/test splits**: 80% for training and 20% for testing.

Table 3. Dataset categories and sizes (the lowest number of comments per demographic category is in bold)

Demographic category	**Categories**	**imbalanced version**	**balanced version**
Age	Under 18 (13-18)	**343709**	343709
	Adult (18-30)	3655357	343709
	Mid-age (30-50)	2890332	343709
	Old (50+)	824243	343709
Gender	Male	18824788	4340722
	Female	**4340722**	4340722

As shown in Table 3, there is a clear dominance of males in the dataset, with female comments representing only about 23% of male comments, which aligns with the findings of [4], who similarly observed male dominance on the Reddit platform. In contrast, our dataset also shows a significant presence of adults and middle-aged individuals, with under 18 representing roughly 9% of adult comments, indicating that these age groups are prominent among Reddit users in the imbalanced version.

3.3. *Author Profiling Models Construction & Evaluation*

After extracting the features from the textual comments, in this phase, we train the ML[§] models including Naive Bayes (NB), LR, RF, and DL[¶] models including Multi Layer Perceptrons (MLP), Convolutional Neural Networks 1 Dimension (CNN1D), and DistilBERT models.

- **NB**: The NB model is a statistical model based on the Bayes theorem of conditional probability. For making a prediction, the model supposes that all features are independent variables to compute the probability of a set of features belonging to a category. However, the independence assumption of NB can be limiting when applied to short social media texts such as Reddit comments, where linguistic features (e.g., words occurrence, syntactic structures) are often correlated. This simplification may cause the model to overlook contextual dependencies between words, reducing its ability to capture nuanced expressions in short texts.
- **LR**: The LR model is a binary classifier that applies a sigmoid function to a linear combination of input features [20]. When dealing with multi-class classification, the model adopts the one-vs-rest strategy, which breaks down the multi-class problem into several binary classification tasks, each distinguishing one class from the rest. Equation (1) represents the probability that a given input $\mathbf{x}$ belongs to the positive class $(y = 1)$ in the logistic regression model.

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}} \tag{1}$$

  Where: $\mathbf{x}$ represents the feature vector of the input data, β_0 is the intercept, $\boldsymbol{\beta}$ is the vector of model coefficients corresponding to the features.

[§] We used the scikit-learn package for ML models implementations

[¶] We used the PyTorch package for constructing and training the DL models

- **RF**: This model is based on the bagging ensemble learning strategy, wherein several shallow decision trees are trained on a subset of data samples and features. The majority vote strategy chooses the final model output by aggregating predictions from all individual trees [21].
- **MLP**: Inspired from the human brain neurons working principles. The model is composed of an input layer, one or more hidden layers of nonlinearly-activating neurons, and an output layer. MLPs are universal function approximators capable of learning complex relationships between input data and desired outputs via backpropagation [22].
- **CNN**: The CNN utilizes convolution layers to extract patterns from input data. Each convolution layer employs a set of learnable filters, defined by their kernel size and weights, which slide across the input data performing convolution operations. This process generates feature maps that capture spatially local patterns. These feature maps can then be further processed by additional convolutional layers, downsampled using pooling layers, or flattened and fed into MLP for final classification or regression. The feature map is computed according to equation (2).

$$F[i,j] = b[j] + \sum_{m=0}^{C_{\text{in}}-1} \sum_{k=0}^{K-1} W[j,m,k]\; X[i\,s + k\,d - p,\, m] \tag{2}$$

  Where: $F[i,j]$ is the output feature value at position i for output channel j; $b[j]$ is the bias for channel j; $W[j,m,k]$ is the weight of the convolution kernel connecting input channel m to output channel j at kernel index k; $X[i\,s + k\,d - p,\, m]$ is the input value at position $i\,s + k\,d - p$ of channel m; K is the kernel size; C_{in} is the number of input channels; and s, p, d denote stride, padding, and dilation, respectively.
- **DistilBERT**: This model is the distilled version of BERT, containing about 66 million parameters, retaining nearly 97% of BERT's language understanding performance while being approximately 40% smaller and 60% faster at inference [23]. In this work we fine-tuned the full model on both tasks.

To determine the best hyper-parameters combination for each model, we performed a hyper-parameters tuning step (see Table 4). Regarding the ML models, we opted for a search grid optimization, while for the DL models, we opted for ten trials with hyper-parameters sampled using employing the Tree-Structured Parzen Estimator [24] using Optuna framework.

In addition to document-level experiments, we included several baseline evaluations to provide a broader comparative framework. These baselines comprise author-level configurations of the DistilBERT model, where multiple comments from the same author are aggregated either directly or through a majority-vote strategy based on document-level predictions. We also incorporated a human classification baseline to approximate human-level performance (using 100 publication per category of each demographic information) and a random classification baseline to represent chance-level predictions. To assess the generalization performance of the trained TF-IDF/LR and DistilBERT models in predicting gender across a different domain, we inferred the labels of the test set from the PAN'13 dataset, which was constructed by collecting blogs from Blogspot [25].

Since we are dealing with a text classification task, we used the Accuracy, Precision, Recall, F1-score and Area Under The Roc Curve (Auc-Roc) metrics [26] .

We mention that all the computations from the dataset storing and annotation to the evaluation were done within MARWAN-HPC[‖], by utilizing computation nodes with 64GB RAM and Intel Xeon(R) Gold 6240R CPU @ 2.40GHz with 48 core and python 3.8, while the DL models training is done within VastAI instances with Nvidia RTX 4090 GPU. Regarding the training of ML models we used the scikit-learn package while for the DL models we used Pytorch. The source code is available on Github[**], and the data is available upon request.

Table 4. Models Hyper-parameters

Model	**Hyperparameters**
NB	alpha in $\{0.25, 0.5, 0.75, 1\}$

LR	regularization = $\{$L2, None$\}$
	max iterations = $\{200, 500, 5000\}$
	C = $\{1, 0.9, 0.75\}$

RF	number of trees = $\{100, 250, 500\}$ max depth = $\{3, 15, 30\}$ criterion = gini

MLP or CNN1D	classification head architecture = $\{$direct out, shallow network, funnel, or long funnel$\}$
	LR = $\{10^{-6}$ to $10^{-2}\}$
	dropout = $\{0$ to $0.4\}$
	epochs = 15
	Loss function = Cross-Entropy
	Optimizer = Adam
	Embeddings = $\{$FastText or Our trained FastText$\}$
	Filters Counts = $\{32, 64, 128\}$
	Kernel size = $\{2, 3, 4\}$
	Pooling = $\{$Max pooling$\}$, Pooling size = $\{2, 3\}$

DistilBERT	classification head architecture = $\{$direct out, shallow network, funnel, or long funnel$\}$
	LR = $\{10^{-6}$ to $10^{-2}\}$
	dropout = $\{0$ to $0.4\}$
	epochs = 15
	Loss function = Cross-Entropy
	Optimizer = Adam

Table 5. Top seven frequent LIWC categories (average by comments) by age class (in percentage)

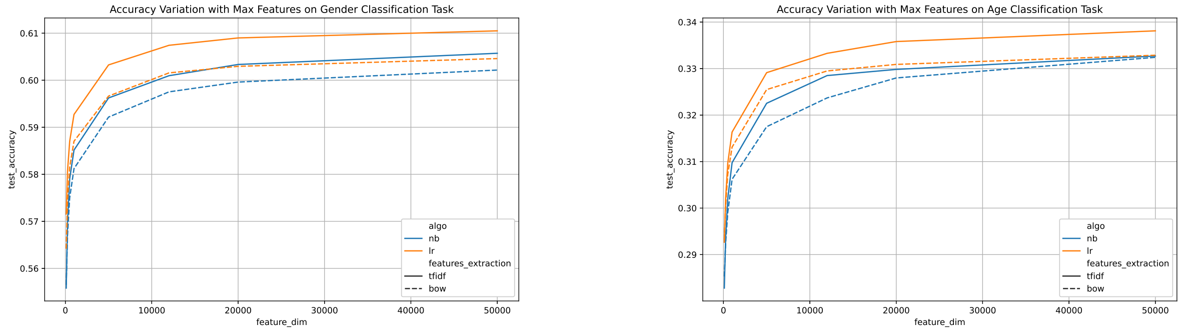
	under 18	18-30	30-50	50+
funct	35.919676	35.879597	35.851102	35.789688
cogmech	11.511431	11.493932	11.433678	11.494826
verb	9.829286	9.805790	9.823365	9.818421
pronoun	8.301021	8.292958	8.301447	8.310523
preps	8.193739	8.148700	8.181808	8.151638
relativ	7.650922	7.672006	7.731151	7.640403
social	6.342213	6.354637	6.321808	6.380742

4. Results & Discussion

In this section, we present and discuss the evaluation results of the trained models. As shown in the line plots in Figure 2 (specifically 2(b) and 2(a)), accuracy varies with different settings for the maximum number of features (based on the top n frequent features) in both the age and gender classification tasks. Notably, the accuracy increases as the maximum number of features rises. In particular, the LR model with TF-IDF feature extraction achieved the

‖https://www.marwan.ma/index.php/en/servicesen/hpc

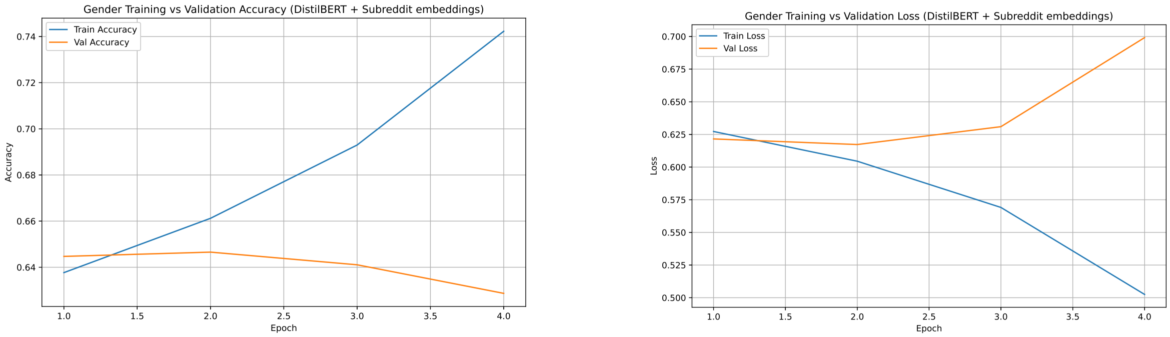
**https://www.github.com/SuicV/demograph_inference_reddit_comments highest accuracy: 0.61 for gender classification and 0.33 for age classification, both with 50,000 maximum features on the test partition. More specifically, the LR model for gender classification surpasses the random classification baseline by 11%, while the age classification model exceeds it by 8%. Compared to the human classification baseline, the same model achieves a higher performance with a 5% improvement in gender classification, whereas for the age classification task, it surpasses the baseline by 3% in terms of F1-score but remains 2% lower in accuracy. These findings highlight a notable success in gender classification, although they also point to areas for improvement in age classification. Moreover, we notice the superiority of TF-IDF features over BOW, LIWC, and POS features used with ML models, as TF-IDF assigns high weights to less frequent tokens and small weights to common tokens. This enables the models to identify differences in word usage between gender and age categories. Additionally, the LIWC and POS features achieve similar performance to the human classification and barely exceed the random classification baselines on both tasks (Tables 6, 9).



(a) Accuracy variation in the gender classification task on the test partition

(b) Accuracy variation in the Age classification task on the test partition

Figure 2. Accuracy variation of NB and LR models with TF-IDF and BOW feature extraction methods in the gender and age classification

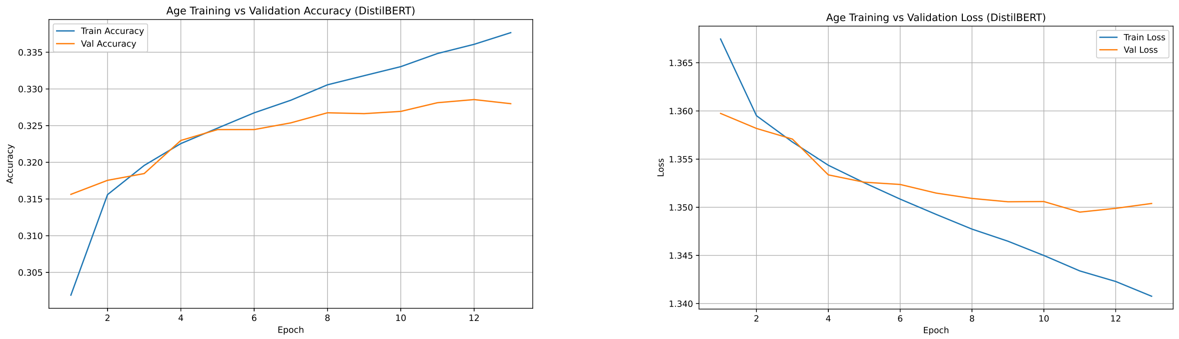


(a) Accuracy variation on the gender classification task on the test partition

(b) Loss variation on the gender classification task on the test partition

Figure 3. Accuracy and loss variation of DistilBERT + Subreddit embeddings model on the gender classification

Figures 3 and 4 illustrate the accuracy and loss variations of the DistilBERT and DistilBERT + Subreddit embeddings models for the gender and age classification tasks, respectively. The DistilBERT + Subreddit embeddings model achieves its optimal loss of 0.61 and an accuracy of 0.65 at the second epoch, after which training stops due to early overfitting. In contrast, for the age classification task, the optimal loss is reached at

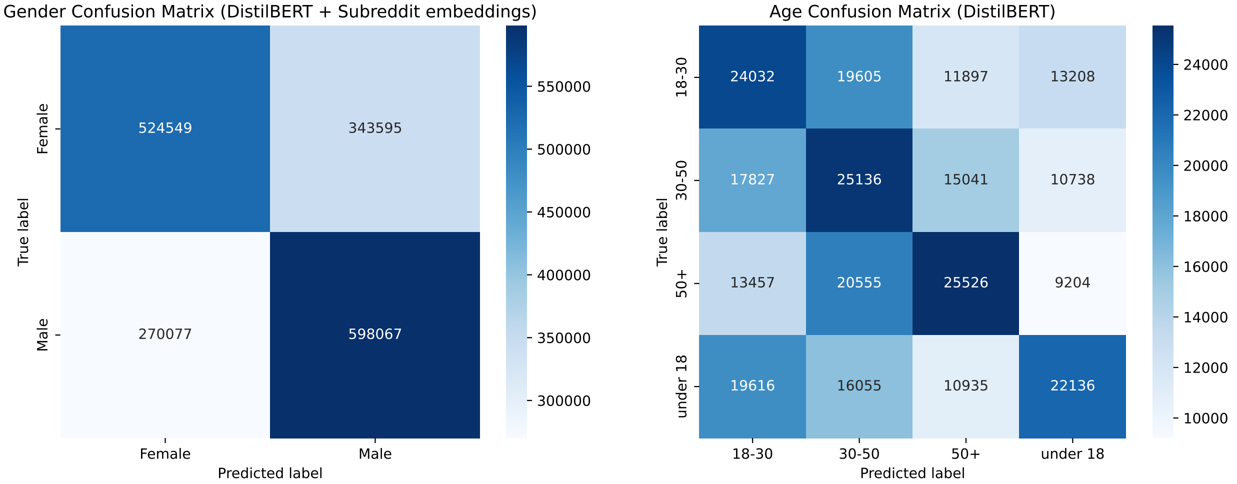


(a) Accuracy variation on the age classification task on the test partition

(b) Loss variation on the age classification task on the test partition

Figure 4. Accuracy and loss variation of DistilBERT model on the age classification

epoch 11, and training is halted at epoch 13 through early stopping. Additionally, the performances of the MLP and CNN1D models are comparable to those of DistilBERT and similar to LR with TF-IDF and BOW, though they remain slightly behind the transformer-based models on both tasks. It is also worth noting that no significant difference was observed between the usage of the original FastText embeddings and our FastText model trained on the Reddit dataset, indicating that domain-specific retraining did not lead to notable performance gains in this context.



(a) Confusion matrix of gender classification on the test partition

(b) Confusion matrix of age classification on the test partition

Figure 5. Confusion matrix of DistilBERT models

As the DistilBERT and DistilBERT + Subreddit embddings achieve the highest accuracy for the age and gender classification tasks, respectively, we present their confusion matrices (Figures 5(b), 5(a)) and per-class performance (Tables 10,7). Regarding the age classification, DistilBERT exhibits moderate and relatively balanced performance across age groups. The model performs best on the "50+" and "under 18" classes, with F1-scores of 0.3863 and 0.3570, respectively, indicating slightly stronger discrimination for the extremes of the age spectrum. In contrast, the "18–30" and "30–50" groups show weaker F1-scores (around 0.33), suggesting that the linguistic patterns of adults and mid-age users are more overlapping and thus harder to distinguish. The confusion matrix further supports this observation, as a large number of instances from these two categories are confused with each other, reflecting shared lexical and stylistic cues. Table 5 also indicates overlaps in word usage across age groups, as only slight differences appear among the most frequent LIWC categories. Similarly, for the gender classification task, the DistilBERT + Subreddit embeddings model achieves the best and most balanced performance across gender groups. The model attains its highest F1-score of 0.66 for the male category, while slightly decreasing to 0.63 for the female category. The confusion matrix (Figure 5(a)) confirms this balance, showing high true positive rates for both classes.

We performed an error analysis, were we analyzed manually 50 publications sampled randomly per category for both tasks. Regarding the gender classification, Female comments were frequently mislabelled when they were short or posted in male-dominated forums like AskReddit and humour subreddits; such comments often lacked relationship words and sometimes used swearing, a stylistic marker more common among male commenters. Male comments tended to be misclassified as female when they contained relationship vocabulary such as "girlfriend" or "mom" or adopted a positive, empathetic tone, features that linguistic studies associate with women's language. On the other hand, for the age classification, most mistakes for under 18 were upgrades to 18–30 or 30–50, typically when youth write in longer, more formal styles or discuss adult-coded topics (politics, work, finance), which blur stylistic cues expected for adolescents. Conversely, 18–30 instances were often shifted to 30–50 (and sometimes 50+) when the text contained family/household or retrospective, nostalgia-tinged narratives that mimic older cohorts. For 30–50, errors split toward 18–30 when slangy tone or gaming/entertainment topics appeared, and toward under 18 when comments were brief and informal.

Generally, in our experiments, the performance of the trained ML and DL models does not exceed 0.65 and 0.37 in accuracy for gender and age classification, respectively. Moreover, the integration of the subreddit from which the comment is published enhanced LR and DistilBERT models prediction performance, especially, the accuracy and Auc-Roc metrics. This limited performance can be attributed to two factors. First, the models lack a global view of each author's writings when operating at the document level, as demonstrated by the improved results of DistilBERT trained at the author level or using the majority vote aggregation of document-level predictions. Second, the relatively low performance of DistilBERT at the author level for age classification is mainly due to the dataset's balancing strategy: while it was balanced at the document level, grouping comments by author resulted in an imbalanced distribution across age categories. In addition, the average comment length in the age dataset is only 38 words, which makes classification particularly challenging at the document-level. We audited group fairness via a disparate-impact style ratio (group metric / best-group metric) with the 80% rule. For gender, all ratios for accuracy, precision, recall, and F1 were ≥ 0.87, indicating no violations though female recall was 0.60 vs 0.69 for males. For age, most ratios satisfied the rule, but precision for the 30–50 class was 0.31 vs a 0.40 reference (ratio 0.78), flagging a disparity.

Regarding the LIME analysis (Table 8) of the TF-IDF/LR model reveals that gender classification is mainly influenced by topic and style related words. Female predictions are driven by terms such as *fat*, *bmi*, *eat*, *tanning*, and pronouns like *she* and *her*, reflecting discussions about appearance and personal care. In contrast, male predictions rely on words related to sports and activities, such as *sports*, *playing*, and *athleticism*, as well as gaming-related vocabulary like *Legends* and *League*, which are associated with discussions of the game *League of Legends*. However, some samples show near-zero attributions, suggesting that the model sometimes bases its decisions on non-discriminative or context-neutral words, highlighting limitations in short or ambiguous comments.

Cross-domain generalization remains a persistent challenge in the AP literature. As shown in Table 11, the performance of our best models, TF-IDF/LR and DistilBERT, trained on Reddit comments, drops to the level of random classification or only marginally surpasses it when evaluated on the Blogs dataset. This degradation is primarily attributed to substantial stylistic and linguistic differences between the two platforms; for instance, Reddit comments are typically short and conversational, whereas blog posts are longer and more structured.

Table 6. Performance of feature extraction methods and models for gender classification (highest score in bold and second best in underline)

Features / Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
BOW / NB	0.60	0.61	0.57	0.59	0.60
BOW / LR	0.60	0.59	0.69	0.64	0.60
TF-IDF / NB	0.61	0.61	0.58	0.60	0.60
TF-IDF / LR	0.61	0.60	0.65	0.63	0.61
TF-IDF + Subreddit embedding / LR	0.63	0.61	0.67	0.64	0.63
LIWC / NB	0.55	0.55	0.59	0.57	0.55
LIWC / LR	0.56	0.55	0.61	0.58	0.56
LIWC / RF	0.56	0.55	0.67	0.61	0.56
POS / NB	0.54	0.55	0.49	0.52	0.54
POS / LR	0.55	0.54	0.55	0.55	0.55
POS / RF	0.56	0.55	0.62	0.58	0.55
AVG FastText / MLP	0.62	0.60	0.68	0.64	0.67
AVG FastText (ours) / MLP	0.61	0.60	0.64	0.64	0.69
AVG FastText / CNN1D	0.60	0.59	0.69	0.61	0.65
AVG FastText (ours) / CNN1D	0.61	0.60	0.65	0.62	0.65
DistilBERT	0.63	0.62	**0.72**	**0.67**	0.69
DistilBERT + Subreddit embedding	**0.65**	**0.64**	0.68	0.66	**0.73**
DistilBERT (Author-Level)	0.74	0.70	**0.85**	0.76	**0.81**
DistilBERT (Author-Level With Majority Vote)	**0.80**	**0.78**	0.79	**0.78**	–
human classification	0.56	0.59	0.57	0.55	–
Random classification baseline	0.50	0.50	0.50	0.50	0.50

Table 7. Per-class metrics for DistilBERT + Subreddit embeddings model on gender classification task.

Class	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Support**
Female	0.65	0.66	0.60	0.63	868,144
Male	0.65	0.64	0.69	0.66	868,144

Table 8. LIME scores per-label.

label	**Scores**
Female	(*fat*, -8.29×10^{-2}), (*weight*, 7.13×10^{-3}), (*bmi*, 6.62×10^{-3}), (*eat*, 3.68×10^{-3}), (*spreadsheet*, 3.13×10^{-3}), *decreases*, 2.97×10^{-3})
Female	(*tanning*, -1.02×10^{-1}), (*orange*, -1.88×10^{-2}), (*she*, 8.81×10^{-3}), (*drips*, 8.08×10^{-3}), (*streaks*, 7.36×10^{-3}), (*dances*, 6.49×10^{-3}), (*wears*, 5.02×10^{-3}), (*foundation*, 4.90×10^{-3}), *her*, 4.80×10^{-3})
Male	(*style*, 3.71×10^{-3}), (*sports*, 2.64×10^{-3}), (*playing*, 1.99×10^{-3}), (*athleticism*, 1.88×10^{-3}), (*their*, 1.79×10^{-3}), (*side*, 1.70×10^{-3}), (*fundamentals*, 1.65×10^{-3}), (*Brazilian*, 1.54×10^{-3})
Male	(*bushes*, 8.86×10^{-32}), (*the*, 8.39×10^{-32}), (*of*, 5.30×10^{-32}), (*can*, 4.71×10^{-32}), (*hide*, -3.12×10^{-32}), (*you*, 2.98×10^{-32}), (*Ha*, -2.67×10^{-32}), (*well*, 1.24×10^{-32}), (*Legends*, 1.04×10^{-32}), (*League*, 7.21×10^{-33})

Table 9. Performance of feature extraction methods and models for age classification (highest score in bold and second best in underline)

Features / Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
BOW / NB	0.33	0.33	0.33	0.33	0.59
BOW / LR	0.33	0.33	0.33	0.33	0.60
TF-IDF / NB	0.33	0.33	0.33	0.33	0.60
TF-IDF / LR	0.33	0.34	0.34	0.33	0.60
TF-IDF + Subreddit embedding / LR	0.36	0.36	0.36	**0.36**	0.63
LIWC / NB	0.29	0.27	0.28	0.25	0.53
LIWC / LR	0.28	0.28	0.29	0.28	0.54
LIWC / RF	0.29	0.29	0.29	0.27	0.54
POS / NB	0.27	0.27	0.27	0.27	0.52
POS / LR	0.27	0.27	0.27	0.26	0.53
POS / RF	0.29	0.29	0.29	0.27	0.54
AVG FastText / MLP	0.33	0.33	0.33	0.32	0.61
AVG FastText (ours) / MLP	0.32	0.31	0.32	0.30	0.60
AVG FastText / CNN1D	0.33	0.33	0.33	0.33	0.60
AVG FastText (ours) / CNN1D	0.31	0.31	0.32	0.30	0.58
DistilBERT	0.35	0.36	0.35	0.35	0.61
DistilBERT + Subreddit embedding	**0.37**	**0.37**	**0.37**	**0.36**	**0.64**
DistilBERT (Author-Level)	**0.36**	0.36	0.36	**0.36**	**0.63**
DistilBERT (Author-Level with Majority Vote)	0.31	**0.39**	**0.44**	0.32	–
human classification	0.35	0.44	0.35	0.30	–
Random classification baseline	0.25	0.25	0.25	0.25	0.50

Table 10. Per-class metrics for DistilBERT model on age classification task

Class	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Support**
18-30	0.65	0.32	0.35	0.33	68,742
30-50	0.64	0.31	0.37	0.33	68,742
50+	0.71	0.40	0.37	0.39	68,742
under 18	0.71	0.40	0.32	0.36	68,742

Table 11. Performance of TF-IDF/LR and DistilBERT on PAN'13 test dataset

Features / Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
TF-IDF / LR	0.49	0.50	0.50	0.49	0.51
DistilBERT	0.52	0.54	0.53	0.49	0.51
Random classification baseline	0.5	0.5	0.5	0.5	0.5

5. Conclusion

5.1. Summary

This paper presents an exploration of the AP task using a large-scale dataset of Reddit comments. By focusing on the document level, we investigated age and gender classification using a combination of ML, DL, and Transformer-based approaches. Our methodology began with dataset annotation using REGEX, followed by text preprocessing and feature extraction with methods such as TF-IDF, BOW, the LIWC dictionary, averaged FastText embeddings, and concatenated Subreddit embeddings.

The experimental results demonstrated that Transformer-based models, particularly DistilBERT and DistilBERT combined with Subreddit embeddings, achieved the best and most balanced performance across both tasks, reaching an accuracy of 0.65 for gender and 0.37 for age classification at the document level, and up to 0.80 at the author level using majority voting. These results confirm that incorporating contextual embeddings and author-level aggregation significantly enhances prediction reliability compared to traditional ML and shallow DL models. Moreover, the study highlights that while gender classification remains more tractable, age prediction continues to pose challenges due to overlapping linguistic patterns across age groups.

5.2. Ethical considerations

As the application of author profiling techniques expands across social media data, it is crucial to embed fairness, privacy, and ethical compliance within model development and deployment. Future work should not only aim for higher predictive accuracy but also ensure that the models operate in an equitable and privacy-preserving manner. Bias mitigation strategies such as reweighting, fair sampling, or adversarial debiasing can be integrated during training to counteract disproportionate representation of demographic groups. The resulting fairness should be systematically assessed using quantitative metrics such as demographic parity, equal opportunity, and equalized odds, providing transparency about how different user groups are affected by model decisions. Equally important are privacy safeguards to prevent potential misuse or re-identification of individuals from textual data. Future research should explore the integration of differential privacy mechanisms, which introduce controlled noise into the training process to protect individual contributions, or federated learning frameworks that enable decentralized training without centralizing user data. In addition, data anonymization procedures (such as hashing usernames, removing personally identifiable information (PII), and discarding nonessential metadata) should be systematically enforced in all stages of data collection and preprocessing. Furthermore, the research should align with international data protection standards, including GDPR and CCPA, by ensuring data minimization, informed consent, and secure storage. Any automated inference about sensitive traits must include a human-in-the-loop review to prevent stigmatization or misinterpretation.

It is essential to emphasize that the proposed models are designed strictly for academic research and social understanding, and should never be used for surveillance, profiling, or decisions impacting individuals. Strengthening adversarial robustness will also help ensure ethical, reliable, and responsible model behavior in future applications.

5.3. Perspectives

There are several promising directions for future research. One key area is to increase the training dataset size to improve model generalization and robustness. A larger dataset would likely enhance the model's ability to capture more complex patterns, especially for the less-represented age groups. Additionally, test larger transformer-based models than DistilBERT, such as BERT base or large [28] and even LLM's such LLaMA or Mistral, which have shown superior performance in various natural language processing tasks. These models could potentially improve classification accuracy at the document level due to their capacity to capture long-range dependencies and nuanced language patterns. Since our best-performing model, DistilBERT, already leverages self-attention mechanisms, future work will also investigate hierarchical pooling strategies to further enhance contextual modeling for short Reddit comments. Moreover, we could address the challenge of dataset imbalance, which can negatively affect model performance. Applying oversampling techniques, such as SMOTE [29], SMOTE-Tomek [30], SMOTE-ENN, or even paraphrasing and synonym replacement, would balance the dataset and reduce the model's tendency to favor dominant classes. Finally, the integration of subreddit embeddings enhanced the performance, we recommend to integration other metadata of the publications as publication timestamp, vote scores, or thread context, and even redesigning LIWC to integrate Reddit linguistics specification such as slang, emoticons. Incorporating these improvements could lead to more robust author profiling models.

REFERENCES

1. D. Roemling, Y. Scherrer, and A. Miletic, *“Explainability of Machine Learning Approaches in Forensic Linguistics: A Case Study in Geolinguistic Authorship Profiling”*, arXiv, Apr. 29, 2024. doi: 10.48550/arXiv.2404.18510.
2. Lanza-Cruz, I., Berlanga, R. & Aramburu, M.J., *Multidimensional Author Profiling for Social Business Intelligence*, Inf Syst Front 26, 195–215, 2024.
3. J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, *“Effects of Age and Gender on Blogging”*, in Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, March 27-29, 2006.
4. M. Thelwall and E. Stuart, *“She’s Reddit: A source of statistically significant gendered interest information?”*, Information Processing & Management, Jul. 2019.
5. F. Hsieh, R. Dias, and I. Paraboni, *“AP from Facebook Corpora”*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018.
6. R. Ramos, G. Neto, B. Silva, D. Monteiro, I. Paraboni, and R. Dias, *“Building a Corpus for Personality-dependent Natural Language Understanding and Generation”*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018.
7. A. Tigunova, P. Mirza, A. Yates, and G. Weikum, *“RedDust: a Large Reusable Dataset of Reddit User Traits”*, vol. Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6118–6126, 2020.
8. A. Tigunova, A. Yates, P. Mirza, and G. Weikum, *“Listening between the Lines: Learning Personal Attributes from Conversations”*, arXiv, Apr. 24, 2019.
9. S. ElSayed and M. Farouk, *“Gender identification for Egyptian Arabic dialect in twitter using deep learning models”*, Egyptian Informatics Journal, Sep. 2020.
10. S. Hussein, M. Farouk, and E. Hemayed, *“Gender identification of egyptian dialect in twitter”*, Egyptian Informatics Journal, Jul. 2019.
11. R. López-Santillán, M. Montes-Y-Gómez, L. C. González-Gurrola, G. Ramírez-Alonso, and O. Prieto-Ordaz, *“Richer Document Embeddings for Author Profiling tasks based on a heuristic search”*, Information Processing & Management, Jul. 2020.
12. *“PAN at CLEF 2015 - Author Profiling.”* Accessed: Apr. 23, 2024. [Online]. Available: https://pan.webis.de/clef15/pan15-web/author-profiling.html
13. *“PAN at CLEF 2014 - Author Profiling.”* Accessed: Apr. 23, 2024. [Online]. Available: https://pan.webis.de/clef14/pan14-web/author-profiling.html
14. S.-H. Cho, D. Kim, H.-C. Kwon, and M. Kim, *“Exploring the potential of large language models for author profiling tasks in digital text forensics”*, Forensic Science International: Digital Investigation, vol. 50, p. 301814, Oct. 2024, doi: 10.1016/j.fsidi.2024.301814.
15. M. Fatima, K. Hasan, S. Anwar, and R. M. A. Nawab, *“Multilingual author profiling on Facebook”*, Information Processing & Management, Jul. 2017.
16. M. A. Ashraf, R. M. A. Nawab, and F. Nie, *“Tran-Switch: A transfer learning approach for sentence level cross-genre author profiling on code-switched English–RomanUrdu Text”*, Information Processing & Management, May 2023.
17. C. Welch, J. K. Kummerfeld, V. Pérez-Rosas, and R. Mihalcea, *“Compositional Demographic Word Embeddings”* in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
18. J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *“The Development and Psychometric Properties of LIWC2007”*.
19. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, *“Advances in Pre-Training Distributed Word Representations”*, in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.
20. Boateng, E. and Abaye, D. *”A Review of the Logistic Regression Model with Emphasis on Medical Research”*. Journal of Data Analysis and Information Processing, 7, 190-207, 2019. doi: 10.4236/jdaip.2019.74012.
21. Rigatti, Steven J. *”Random forest”*. Journal of Insurance Medicine 47.1 (2017): 31-39.
22. A. Shrestha and A. Mahmood, *”Review of Deep Learning Algorithms and Architectures”*, in IEEE Access, vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.
23. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *“DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”*, 2020, arXiv: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.
24. S. Watanabe, *”Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance”* May 26, 2023, arXiv: arXiv:2304.11127. doi: 10.48550/arXiv.2304.11127.
25. F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. *”Overview of the Author Profiling Task at PAN 2013”*. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, September 2013. CEUR-WS.org. ISBN 978-88-904810-3-1. ISSN 2038-4963.
26. Z. D. Vujovic, *“Classification Model Evaluation Metrics”*, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 6, Art. no. 6, 30 2021, doi: 10.14569/IJACSA.2021.0120670.
27. A. Vaswani et al., “Attention Is All You Need.”, arXiv, Jun. 12, 2017. doi: 10.48550/arXiv.1706.03762.
28. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”*, arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.
29. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *“SMOTE: Synthetic Minority Over-sampling Technique”*, jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
30. M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, *“Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data”*, in 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), May 2016, pp. 225–228. doi: 10.1109/ICOACS.2016.7563084.