# Experimental Analysis of the Relevance of Features and Effects on Gender Classification Models for Social Media Author Profiling

Paloma Piot–Perez-Abadin[a], Patricia Martin–Rodilla[b] and Javier Parapar[c]
*IRLab, CITIC Research Centre, Universidade de Coruña, A Coruña, Spain*

Keywords: Gender Classification, Author Profiling, Feature Relevance, Social Media.

Abstract: Automatic user profiling from social networks has become a popular task due to its commercial applications (targeted advertising, market studies...). Automatic profiling models infer demographic characteristics of social network users from their generated content or interactions. Users' demographic information is also precious for more social worrying tasks such as automatic early detection of mental disorders. For this type of users' analysis tasks, it has been shown that the way how they use language is an important indicator which contributes to the effectiveness of the models. Therefore, we also consider that for identifying aspects such as gender, age or user's origin, it is interesting to consider the use of the language both from psycho-linguistic and semantic features. A good selection of features will be vital for the performance of retrieval, classification, and decision-making software systems. In this paper, we will address gender classification as a part of the automatic profiling task. We show an experimental analysis of the performance of existing gender classification models based on external corpus and baselines for automatic profiling. We analyse in-depth the influence of the linguistic features in the classification accuracy of the model. After that analysis, we have put together a feature set for gender classification models in social networks with an accuracy performance above existing baselines.

## 1 INTRODUCTION

Automatic author profiling is a research area that has gained some relevance in recent years focused on inferring social-demographic information about the author or user of a certain application or software service (Álvarez-Carmona et al., 2016). The growing number of studies in author profiling is mainly explained by the large number of possible applications in strategic sectors such as security, marketing, forensic, e-commerce, fake profiles identification, etc. (Rangel et al., 2014).

Recent author profiling efforts include the development of shared tasks and corpora for evaluating author profiling, especially taking written texts by the user as relevant information for the demographic profile construction. Resultant author profiling models show that the language used in social network publications is a very relevant demographic indicator, identifying aspects such as gender, age or user's origin from psycho-linguistic and semantic features. These features play an important role in retrieval, classification, and decision-making software systems.

However, some authors have begun to study this area from a critical perspective, demanding more "corpora benchmarks to develop and evaluate techniques for author profiling" (Fatima et al., 2017) and focusing on the classification model design as an extremely time-consuming task which requires some reduction efforts and comparative analysis. As an important part of the existing corpora and shared tasks for author profiling, gender classification is a crucial part of the demographic profile. Current classification models present a high number of features about the user's behaviour and their linguistic style in written texts and similar semantic variables that increase the complexity and the time consumed in the design and execution of gender classification models.

This paper shows an experimental analysis of the performance of existing gender classification models. We use external corpora and baselines from well-known author profiling shared tasks for our analysis. The results allowed us to identify linguistic and semantic features with special relevance in gender classification models from social networks, obtaining a

[a] https://orcid.org/0000-0002-7069-3389
[b] https://orcid.org/0000-0002-1540-883X
[c] https://orcid.org/0000-0002-5997-8252



Piot–Perez-Abadin, P., Martin–Rodilla, P. and Parapar, J.
Experimental Analysis of the Relevance of Features and Effects on Gender Classification Models for Social Media Author Profiling.
DOI: 10.5220/0010431901030113
In *Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2021)*, pages 103-113
ISBN: 978-989-758-508-1

feature-combined model for gender classification that improves existing baselines in accuracy performance.

The paper is structured as follows: Section 2 introduces the necessary background for the paper, including works and real applications on author profiling, specific cases in gender classification from social networks and needs in terms of features studies and their linguistic basis. Section 3 explains the experimental analysis design including external data and baselines used, and the experimental workflow carried out. Section 4 details the experiment results and the final model achieved. Section 5 discusses the final results obtained, presenting some conclusions about application possibilities and outlining future work.

## 2 BACKGROUND

It is common to include automatic profiling software in marketing analysis and decision-making processes, where certain companies and services are interested in automatic profiling algorithms. The main goal is knowing their current users or their potential market for redirecting their advertising campaigns, as well as evaluating the opinions of their users about products or services.

In a different large group of applications, we can find some more user-centred researches where automatic profiling algorithms allow a forensic analysis at a behavioural and psycho-linguistic level of the author or user of a certain application or software service, such as blogs (Mukherjee and Liu, 2010), social networks (Alowibdi et al., 2013; Peersman et al., 2011), etc. This analysis is being used successfully to detect early risk on the internet of certain behaviours (e.g. cyberbullying (Dadvar et al., 2012), hate speech (Chopra et al., 2020), etc.) and certain mental disorders (depression (Losada et al., 2019; Losada et al., 2018), bipolar disorder (Sekulic et al., 2018), anorexia (Losada et al., 2017) etc.).

It is also possible to distinguish different main approaches to the problem according to the primary source of information used. First of all, we find in the literature many studies and applications already in production that use images or audiovisual material as a primary source for automatic profiling (XueMing Leng and YiDing Wang, 2008; Makinen and Raisamo, 2008). Images or videos that can be shared by a user on a social network, blog or web, or being part of a more private repository as confidential information, such as medical repositories, etc. In this type of approach, the classification algorithms, using for example Support vector machines (SVM) (Moghaddam and Ming-Hsuan Yang, 2000) or Convolutional Neural Network (CNN) have offered successful results (Levi and Hassner, 2015).

Another possible approach takes as source information the behaviour of the user, their movements and decisions using software systems. This approach is common in author profiling from social networks or in applications linked to the consumption of online services. In general, the behaviour-based approach tends to include behavioural variables (pages visited, links, connection times, purchases made, colour-based studies, etc.) (Alowibdi et al., 2013; Alowibdi et al., 2013; Peersman et al., 2011) as information for author profiling (e.g. age or gender classification), with results around 0.60 accuracy. To improve these results, behavioural variables are usually combined with semantic and psycho-linguistic variables based on the analysis of the user's textual comments (posts on social networks, reviews of services, etc.), requiring feature identification with a high linguistic base. Recent author profiling shared tasks and studies have explored lexical, grammatical or discursive components (Miller et al., ; Ortega-Mendoza et al., 2016) as features for author profiling, also in multilingual environments (Fatima et al., 2017).

Thus, the natural language used in the publications of the social network is relevant for the automatic profiling task. Most recent results offer classification models for certain aspects of author profiling (mainly age and gender) with accuracy over 0.70 (Koppel et al., 2002). Vasilev (Vasilev, 2018) makes a complete review of these systems in the specific case of Reddit and, using controlled subreddits as sources, achieves an accuracy close to 0.85, showing the potential of the hybrid behavioural and linguistic combination for automatic profiling from social networks.

However, all these success cases require a high number of features in their classification models to obtain above 0.70 of accuracy rates, especially semantic and psycho-linguistic features. Some authors have already warned of the excessive number of features involved in classification systems for automatic profiling, which makes the task of designing classification models for these systems a very time-consuming task, and recommend specific studies to reduce the number of features in the models (Alowibdi et al., 2013).

Taking into account the potential of automatic profiling systems based on semantic and linguistic features and the need of experimental studies on existing classification models for reducing their complexity, this paper focuses on gender classification as a specific feature of automatic profiling. Gender is used as a differential factor in the treatment and detection of early risk signs in mental disorders (Aggarwal et al.,

2020), which makes gender crucial information on the demographic profile of a user for these applications in social networks.

The following sections show the design of experiments carried out to study the features involved in gender classification in several author profiling tasks, determining their relevance in the classification models created. Subsequently, we propose a gender classification model with a reduction of features based on the relevance found, achieving an improvement in terms of accuracy.

# 3 DESIGN OF EXPERIMENTS

## 3.1 Workflow

Our study presents two phases. Firstly, a phase where a performance analysis of the most used gender classification models is carried out on the selected external datasets, tracking each included feature on each classification model in terms of the relevance and effects of their inclusion in the model analyzed. We will use each model accuracy (Metz, 1978) as a performance metric, measuring the accuracy of the models presented. This phase gives us results in two directions: 1) which classification algorithms obtain the best results for gender and 2) which of the semantic and linguistic features incorporated have greater relevance in the model, that is, which of them contribute more to achieve the accuracy reported.

Once we have the most relevant features, a second phase of experiments is carried out, building a model that includes most relevant features and using gender classification algorithms with better results. This model serves as a base model for gender classification in automatic profiling with a significant reduction in features.

The experiments carried out in both phases share the design workflow, which follows a classic process of experimentation in classification algorithms, with two main processes: the training process and the test process. Both train and test processes consist of a preprocessing step, where we convert the raw data into a data frame, and a feature engineering stage, where we obtain the features from the corpus.

The training process continues by splitting the dataset into a train and a test subset to train our models. We apply cross-validation and, the output will be the resultant classification model.

The test phase takes the classification models to predict the unseen data and gives us the accuracy of the models. The workflow followed in each experiment carried out is detailed in Figure 1).

## 3.2 Datasets

Most well-known efforts on author profiling are PAN[1] initiatives, with shared tasks editions between 2013 and 2020. We have selected two different datasets which include gender information from PAN as external corpora for our experimental analysis. Specifically, we have selected from the competition the "PAN Author Profiling 2019" and the "PAN Celebrity Profiling 2019" datasets, both in English. We have selected these datasets because they are the most recent available datasets at the time of our experiments' design phase (the 2020 edition data was not known until this 2020's fall). All PAN datasets are available in their website[2].

The first dataset is divided into two parts: a training dataset and a test dataset. Both are composed of bots and humans users, because the main goal of the task was to identify if the author on Twitter information is a human or a bot user, and, in case of human, it is necessary to infer the user's gender. Hence, we have filtered out the bots for our repository. We have used each dataset in its respective phase of our experiments.

The second dataset presents only training part with Twitter information. In this case, the goal of the shared predicts celebrity traits from the teaser history. In terms of celebrity gender, we had to filter out only 18 "nonbinary" users (as we modelled gender only in male/female situations for our applications). We have used the second dataset only the in training phases of our experiments.

Next subsection details in depth both datasets in terms of volume and internal characteristics.

### 3.2.1 "PAN Author Profiling 2019" Dataset

"PAN Author Profiling 2019" dataset (pan, ) comes from the "Bots and Gender Profiling 2019" PAN shared task. The goal of this task was: "Given a Twitter feed, determine whether its author is a bot or a human. In the case of human, identify her/his gender".

We decided to use this dataset as it is a balanced collection containing social media texts without any cleaning action and the interactive social media nature of the data is also more suitable for our objectives. Furthermore, this dataset has streams of text with temporal continuity in the writings. This allows the appearance of linguistic phenomena in the contributions of the writers that are likely to be relevant as features for our gender classification study.

[1] https://pan.webis.de/

[2] https://pan.webis.de/shared-tasks.html

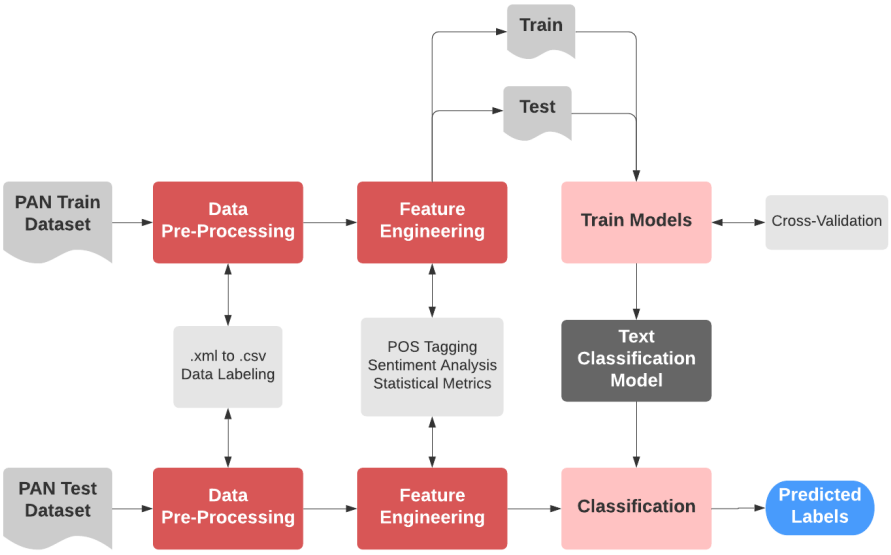


Figure 1: Gender classification task workflow.

Table 1: "Author Profiling 2019" dataset.

| | author_id | gender |
|---|---|---|
| 0 | ccbe6914a203b899882 | male |
| 1 | a3b93437a32dba31def | male |
| 2 | a1655b4b89e7f4a76a9 | male |
| 3 | de3eee10fbac25fe396 | male |
| 4 | 2a61915c1cd27b842ee | male |
| ... | ... | ... |
| 2055 | f92806b515385388c83 | female |
| 2056 | a820cb38384e19a3043 | female |
| 2057 | f17345aeea69b649063 | female |
| 2058 | f334e25ccf9a18f1eb2 | female |
| 2059 | b2eb427fb56beace062 | female |

Each user is represented in the dataset as a `.xml` file. Each file contains an `author` tag, which includes a `documents` tag holding a list of 100 `document` symbolizing one Tweet each. This dataset has a `truth.txt` master file containing the gender label for each author. See Table 1 for entrance examples.

After preprocessing the dataset, we have ended up having a balanced dataset with 2060 users: 1030 female users and 1030 male users. (See Table 2).

Table 2: "Author Profiling 2019" dataset users by gender.

| | total |
|---|---|
| male | 1030 |
| female | 1030 |

#### 3.2.2 "PAN Celebrity Profiling 2019" Dataset

"PAN Celebrity Profiling 2019" dataset[3] comes from the "Celebrity Profiling 2019" PAN task. The goal of this task was: "Given a celebrity's Twitter feed, determine its owner's age, fame, gender, and occupation".

We have decided to use this dataset for the same reasons as described in "PAN Author Profiling 2019" dataset and, besides it contains plenty of user-profiles and an average of 2000 Tweets per user. Thus it will complement the author's competition dataset properly.

The available data consist of two files: the `feed` file containing the author id and a list of all Tweets for each user and the `labels` file holding the author id and the value for each trait. The traits are fame, occupation, birth year, and gender. In the present study, we have disregarded all but gender. See Table 3 for entrance examples.

After preprocessing the dataset, we have winded up obtaining an unbalanced dataset, with twice more male users than female users. (See Table 4).

Due to the complexity of the task and the final goals of our study, these have been the datasets that best suit our needs. Therefore, we have decided to use these in our experiments. Moreover, these sets would let us be able to compare our results with a clear baseline.

### 3.3 Data Preprocessing

Each combination of gender classification algorithm + classification model features reproduces the same

[3]https://pan.webis.de/data.html

Table 3: "PAN Celebrity Profiling 2019" dataset.

| | author_id | gender |
|---|---|---|
| 0 | 3849 | male |
| 1 | 957 | female |
| 2 | 14388 | female |
| 3 | 7446 | male |
| 4 | 1107 | female |
| ... | ... | ... |
| 14494 | 33530 | female |
| 14495 | 29315 | male |
| 14496 | 36954 | male |
| 14497 | 4554 | male |
| 14498 | 4512 | male |

Table 4: "PAN Celebrity Profiling 2019" dataset users per gender.

| | total |
|---|---|
| male | 10409 |
| female | 4072 |
| nonbinary | 18 |

workflow detailed in Figure 1. First, and before performing feature engineering on the data, we have transformed the different documents to have a homogeneous set. For the "PAN Author Profiling 2019" we have joined each `.xml` into one file and have converted it into a `.csv` file. In the case of "PAN Celebrity Profiling 2019", we have merged both `.json` file and have transformed it into a `.csv` file.

The `.csv` file consists of an `id` column, a `text` column enclosing all messages (our corpus), and a `gender` column (what we want to predict).

We have decided not to carry out any additional preprocessing steps. Applying stemming or lemming would suppose a loss of potentially relevant information for the classification task, just like removing stopwords and special characters since the corpus will be significantly abridged, and therefore entailing a loss of content and precision.

## 3.4 Feature Engineering

The main idea behind feature engineering is using domain knowledge to obtain features from the corpus. To find a characteristic pattern between diverse authors, we have used these features. The first goal of this study is to analyze what kind of features present current models of gender classification, analyzing their relevance and effects on the model. For this purpose, we have divided the found features and some extra features added by us into three groups based on the intrinsic nature of the information involved:

1. Sociolinguistic features
2. Sentiment Analysis features
3. Topic modelling features

### 3.4.1 Sociolinguistic Features

Sociolinguistics is the study of the effect of any aspect of society on the way language is used (Rajend et al., 2009; Coates, 2015). In sociolinguistics, gender refers to sexual identity concerning culture and society. How words are used can both reflect and reinforce social attitudes toward gender. From this definition, we have calculated how many times a peculiar stylistic feature appears in the text concerning this category. This approach will help us, for instance, to find a common generalized lexicon shared by males and another for females, or to infer grammatical or discursive structures uses with difference by gender.

In particular, some of the features in this group that we have found in previous models or have included in our experiments refer to:

- Emojis features, which are related to the lexical derivation and new word formation and with the semantics and neurolinguistics implications of emotion and symbols.
- Punctuation marks, which are related to syntaxis and speech analysis, like word and text length.
- Features as repeated alphabets, readability, and cosine similarity as pragmatic features, among URLs and hashtags, which at the same time are personal-temporal-space references.
- Self-referentiality is a feature that is usually added to stop words, and thus it is not recorded, but it has a powerful sociolinguistic effect on the classification task.
- Part-of-Speech (POS) information in form of POS Tags are sociolinguistic traits taking into account the syntax but also related to speech and discourse analysis.
- Readability is a metric that indicates how difficult a text in English is to understand. It is defined by Flesch-Kincaid reading ease, where a higher score indicates that the passage is easier to read (Ease, 2009). The following formula is used to calculate this score:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

A summary on the sociolinguistic features analyzed is shown in Table 5. As technical level, we have used

regular expressions, *pandas*, and libraries like *scikit-learn*[4], *NLTK*[5], *spaCy*[6], *pyphen*[7] and *Emoji*[8], to extract and perform this features analysis from datasets information.

#### 3.4.2 Sentiment Analysis

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, emotions, etc. from written language (Liu, 2010; Liu, 2012). In this process, we usually try to determine whether a piece of writing is positive, negative, or neutral.

We have made use of *NLTK* sentiment analysis analyzer to extract the compound and neutral scores for each document in the corpus. Sentiment analysis helps us to understand the author's experiences and can be a pattern differentiated between males and females. Thus, sentiment analysis information constitutes also a feature to take into account in gender classification models.

#### 3.4.3 Topic Modelling

Latent Dirichlet Allocation (LDA) is a topic model proposed by David Blei et al. (Blei et al., 2003) used to classify text in a document referring to a particular topic. It is an unsupervised generative statistical model which builds a topic per document model and words per topic model, modelled as Dirichlet distributions.

LDA is commonly used for automatically extracting and finding hidden patterns among the corpus. This feature can help us modelling relations between topics for each gender category.

We have followed this approach for topic modelling. We have executed this solution in order to fetch the twenty most significant topics, defined with twenty words each, and constituting also (twenty) features in our gender classification study. These features are represented by the numbers from 1 to 20. In Figure 2 the features tagged with the numbers from 1 to 20 represent these topics.

### 3.5 Classification Algorithms and Experimental Configurations

We have examined different classifiers and, considering gender classification as a binary classification task, we have taken into account different approaches making use of the following algorithms. To find the best performance, we have performed a grid search to hyper-parameter tuning experiments. Finally, we have run our experiments with the following algorithms and configurations:

- *Random Forest*, the default configuration of sklearn implementation of this algorithm, except the estimators (500). As this is a sophisticated task, we decided to try Random Forest to be able to learn a non-linear decision boundary and try to achieve higher accuracy scores than with a linear-based algorithm (Kirasich et al., 2018).
- *Adaptive Boosting*, base estimator Decision Tree, 500 estimators, algorithm SAMME, maximum depth 9. The main idea behind boosting algorithms is to train predictors sequentially, each one trying to correct the errors of its predecessor in each iteration so that the next classifier is built based on the classification error of the previous one. Numerous PAN competition participants have tried out AdaBoost on the bots recognition task, so we have decided to attempt AdaBoost in the gender classification task (Bacciu et al., 2019).
- *LightGBM*, 500 iterations, maximum depth 7, learning rate 0.15, gbdt boosting, metric binary logloss, min data in leaf 600, bagging fraction 0.8, feature fraction 0.8. We have decided to use LightGBM because it is a gradient boosting framework that uses tree-based learning algorithms. It focuses on the accuracy of results and it is one of the algorithms that is leading classification competitions (Ke et al., 2017).

## 4 EXPERIMENTS

We have carried out combinations of experiments to see which is the most effective classifier for the gender profiling task, evaluating its precision to find out which is the most suitable.

We have trained the model making use of the "PAN Author Profiling 2019" train dataset and "PAN Celebrity Profiling 2019" dataset. We made our classification results based on the accuracy of the model on the "PAN Author Profiling 2019" test dataset.

In Table 6 we show our experiments results. Each column represents the dataset used for training each experiment, while each row represent the features combination and the algorithm used. We have represented "PAN Author Profiling 2019" train dataset as "Author" and "PAN Celebrity Profiling 2019" as "Celebrity". The combination of both is "Author + Celebrity". Moreover, the validation of our models

[4]https://scikit-learn.org/stable/
[5]https://www.nltk.org/
[6]https://spacy.io/
[7]https://pyphen.org/
[8]https://github.com/carpedm20/emoji/

Table 5: Sociolinguistic features described.

| Feature Names | Feature Description |
|---|---|
| Emojis use | Emojis use ratio per user documents |
| Separator and special characters | Blanks/brackets/ampersand/underscore ratio per user documents |
| Punctuation marks | Question/exclamation/punctuation marks ratio per user documents |
| URLs, hashtag | Separately URLs and hashtags ratio per user |
| Tokens | Words ratio per user document |
| Words length, text length | Mean word length and text length |
| POS Tags | Part-of-speech tagging: ratio per user documents |
| Repeated alphabets | Repeated alphabets ratio per user documents |
| Self-referentiality | Ratio of sentences refering to ifself |
| Readability | Metric for the ease with which a reader can understand a written text |
| Cosine similarity | Measure of similarity between two documents (in all user documents) |

has taken place with "PAN Celebrity Profiling 2019" dataset, so the accuracy is regarding this dataset.

Regarding featuring combination, we have tested the models: 1) with initial sociolinguistic features without topic information, 2) with all initial features + adding topic modelling information, 3) eliminating the features with zero relevance on the data -in Figure 2, the ones with zero coefficient-, 4) removing the less important features -in Figure 2, removing the ones with a coefficient lower than 20- and 5) top half important features -in Figure 2, keeping the ones with a coefficient higher than 50-.

Regarding classification algorithms, we have trained Random Forest; Ada Boost with Decision Tree as a base estimator and LightGBM; with the mentioned parameter configuration. Accuracy results on 1, 2, 4 and 5 feature configuration (combination 3 offered the same accuracy as 2) are shown in Table 6.

It is important to highlight that the experiments carried out also allow us to perform an analysis on each specific feature, its relevance and effects on each model. Thus, Figure 2 shows the relevance of each feature in the classification model with higher accuracy (LightGBM). The graph is generated by LightGBM by using the function `feature_importance()`, and the importance type is calculated with "split", which means that the result contains the number of times a feature is used to split the data across all trees.

We can see that love emojis, exclamation marks, affection emojis, and articles together with one of the topic models are the most important features in the classification model studied. This feature relevance metric shows the number of times the feature is used in a model, the higher, the more important.

Other features like cosine similarity, self-referentiality, readability, numbers, interjections, and adjectives are among the half more relevant features.

Determiners, monkey emojis, word ratio, and word length mean, has a low impact, but still, their contribution is relevant as the classifier gets worse if we remove them.

On the other hand, we can find some features that are good candidates for removing: square brackets, lines, and coordinating conjunctions have no real impact on our classification models. Thus, removing them by evaluating the models with zero coefficient on them makes the models maintain the same accuracy. These features seem to have small relevance in the classifiers models, although in some cases, removing those it can bring about a decrease in the model.

Comparing the results, the best approach we have achieved was using LightGBM learning algorithm with LDA topics and keeping all features. We got an accuracy of **0.7735**, and we can compare our result with the classification accuracy in "PAN Author Profiling 2019" task, as we have validated our models with the test dataset given by PAN. As we decided to study the primary linguistic features, our accuracy results are not among the top of PAN competition, but still, we have demonstrated that this approach gives a good result (almost 4 out of 5 are well classified), and combining this approach with word and char n-grams can result on an excellent classifier.

We can also remark that the combination of both available datasets gives a better result because the model can generalize more advantageously. Therefore the outcome in unseen data is slightly higher than the other experiments.

## 5 DISCUSSION AND CONCLUSIONS

First of all, it is important to highlight that, due to merging both author and celebrity datasets results in a higher accuracy model, it is possible that merging similar sources for obtaining a higher volume of training data, we could improve the presented results.

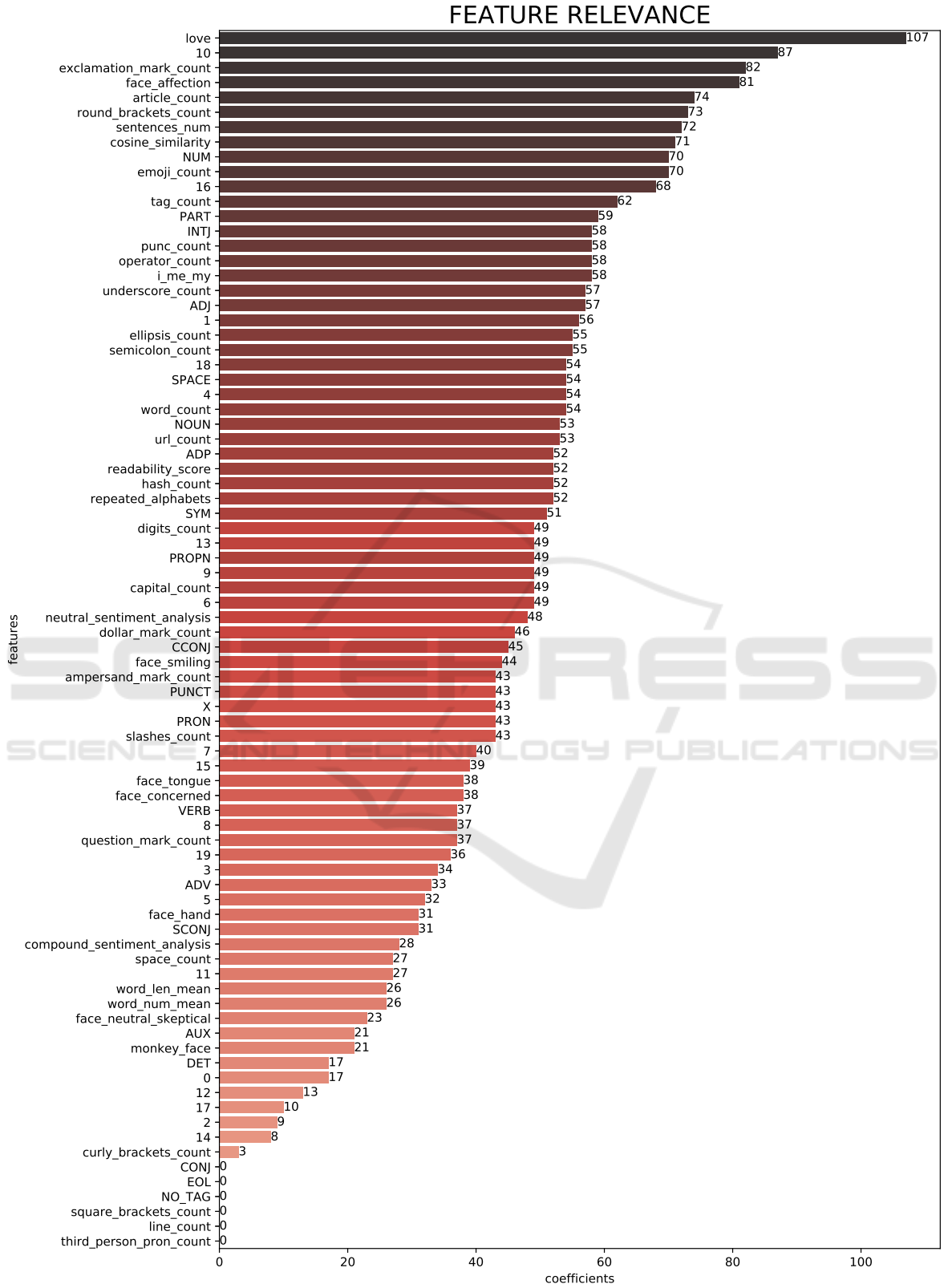


Figure 2: Results on feature relevance.

Table 6: Gender classification accuracy.

| Model + features | Author | Celebrity | Author + Celebrity |
|---|---|---|---|
| **All features and topics** | | | |
| RandomForestClassifier | 0.6030 | 0.5470 | 0.7174 |
| AdaBoostClassifier | 0.5538 | 0.5114 | 0.7470 |
| LightGBM | 0.5818 | 0.6121 | **0.7735** |
| **All features without topics** | | | |
| RandomForestClassifier | 0.6621 | 0.5174 | 0.6780 |
| AdaBoostClassifier | 0.6523 | 0.5197 | 0.6780 |
| LightGBM | 0.6598 | 0.5795 | 0.7152 |
| **Without less important features** | | | |
| RandomForestClassifier | 0.6720 | 0.5303 | 0.6985 |
| AdaBoostClassifier | 0.6598 | 0.5083 | 0.7258 |
| LightGBM | 0.6803 | 0.5068 | **0.7561** |
| **Top half important features** | | | |
| RandomForestClassifier | 0.6530 | 0.5523 | 0.6583 |
| AdaBoostClassifier | 0.6545 | 0.5871 | 0.6795 |
| LightGBM | 0.6538 | 0.5985 | 0.7008 |

So, our study and the resultant models present some improvements in accuracy comparing with some reported results on gender classification PAN tasks[9]. This study has shown how linguistic features such as sentiment analysis or topic modelling play an important role in gender classification for author profiling.

Another interesting effect shown in the experiments is that, although there are almost non-relevant features in the models, removing some features with very low relevance coefficients lead to a decrease in the accuracy: this means that every characteristic we have calculated denotes effect in some way (i.e. positive or negative effects on the gender classification model). This situation somehow justifies the previous trend of adding features to the classification models until reaching the enormous number of features reported in recent works. However, we must take into account the effective cost of the design and execution of these models with so many features. Thus, this study presents quantitative results on features relevance that allow us to analyze the impact of removing features in some classification combinations. As can be seen in Table 6, the accuracies obtained 1) for the best algorithm -LightGBM- in the case of the complete model (with all the features) and 2) for that algorithm -LightGBM- in the case of the efficient model (without less important features) are not so far apart.

These results open space for discussing in which cases and for which software systems a slightly higher accuracy is necessary in the case of gender classification but with a much higher design and execution time-consuming models (due to the large number of features), or in which cases we can apply the efficient model with the lowest number of features without compromising the gender classification of the software system. Although there is still work to be done to improve accuracy for models with less number of features, the savings in terms of design and execution effort can outweigh many applications.

As immediate future steps, and following the results detailed here, it is necessary to generalize even more the original datasets used as sources of information. As we obtain better results on mixing datasets, we plan to run the same battery of experiments including other datasets with gender information to increase the generalization possibilities of our study.

We also plan in the future to perform similar studies on specific aspects of author profiling tasks, such as age or socio-economic variables (income level, etc.) classification and inferring. These studies present different challenges compared to the case of gender classification, mainly due to the non-binary nature of the classification problem. Due to this, the number of possible approaches and possible features for these models is even greater than for binary classification problems, which makes these studies relevant to alleviate the problem of the large number of features in the models. For these challenges, we could use even a more linguistic-based approach, such as n-grams or n-chars models. Including these approaches

[9]https://pan.webis.de/data.html

could help us to achieve better results. Also, trying a word embedding approach may result in an improvement of the models, but we have decided to focus on studying the importance of the primary features.

Finally, we plan to test the classification models with the best performance (best accuracy with a fewer number of features) both for gender and for the rest of the automatic profiling aspects in real applications that are already doing author profiling. These tests will allow us to evaluate efficient models versus those that are already in use (with a higher number of features), analyzing whether it is worth putting less time-consuming models into production (with fewer features) achieving similar accuracy results. Thus, we could figure out which type of application it is appropriate employing the efficient model or on the contrary, in which applications it is better to opt for the time-consuming solution.

## ACKNOWLEDGEMENTS

This work was supported by projects RTI2018-093336-B-C21, RTI2018-093336-B-C22 (Ministerio de Ciencia e Innvovación & ERDF) and the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G/01, ED431B 2019/03) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT of the University of A Coruña as a Research Center of the Galician University System.

## REFERENCES

Aggarwal, J., Rabinovich, E., and Stevenson, S. (2020). Exploration of gender differences in covid-19 discourse on reddit.

Alowibdi, J. S., Buy, U. A., and Yu, P. (2013). Empirical evaluation of profile characteristics for gender classification on twitter. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 365–369.

Alowibdi, J. S., Buy, U. A., and Yu, P. (2013). Language independent gender classification on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, page 739–743, New York, NY, USA. Association for Computing Machinery.

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., and Meza, I. (2016). Evaluating topic-based representations for author profiling in social media. In Montes y Gómez, M., Escalante, H. J., Segura, A., and Murillo, J. d. D., editors, *Advances in Artificial Intelligence - IBERAMIA 2016*, pages 151–162, Cham. Springer International Publishing.

Bacciu, A., Morgia, M. L., Mei, A., Nemmi, E. N., Neri, V., and Stefa, J. (2019). Bot and gender detection of twitter accounts using distortion and LSA notebook for PAN at CLEF 2019. *CEUR Workshop Proceedings*, 2380(July).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3(4–5), 993–1022*.

Chopra, S., Sawhney, R., Mathur, P., and Shah, R. R. (2020). Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 386–393.

Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language, third edition (pp. 1–245)*. Taylor and Francis.

Dadvar, M., Jong, F. d., Ordelman, R., and Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Ease, F. R. (2009). Flesch–Kincaid readability test.

Fatima, M., Hasan, K., Anwar, S., and Nawab, R. M. A. (2017). Multilingual author profiling on facebook. *Inf. Process. Manage.*, 53(4):886–904.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, Inc.

Kirasich, K., Smith, T. ., and Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Liu, B. (2010). Sentiment analysis and subjectivity. In *In Handbook of Natural Language Processing, Second Edition (pp. 627–666). CRC Press*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5(1), 1–184*.

Losada, D. E., Crestani, F., and Parapar, J. (2017). erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In Jones, G. J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Inter-*

*action*, pages 346–360, Cham. Springer International Publishing.

Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of erisk: early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 343–361. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D. E., Heinatz Bürki, G., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, Cham. Springer International Publishing.

Makinen, E. and Raisamo, R. (2008). Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. WB Saunders.

Miller, Z., Dickinson, B., and Hu, W. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science (2012) 02(04) 143-148.*

Moghaddam, B. and Ming-Hsuan Yang (2000). Gender classification with support vector machines. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 306–311.

Mukherjee, A. and Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA. Association for Computational Linguistics.

Ortega-Mendoza, R. M., Franco-Arcega, A., López-Monroy, A. P., and Montes-y Gómez, M. (2016). I, me, mine: The role of personal phrases in author profiling. In Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 110–122, Cham. Springer International Publishing.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, SMUC '11, page 37–44, New York, NY, USA. Association for Computing Machinery.

Rajend, M., Swann, J., Deumert, A., and Leap, W. (2009). *Introducing sociolinguistics*. Edinburgh University Press.

Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W., et al. (2014). Overview of the 2nd author profiling task at pan 2014. In *CEUR Workshop Proceedings*, volume 1180, pages 898–927. CEUR Workshop Proceedings.

Sekulic, I., Gjurković, M., and Šnajder, J. (2018). Not just depressed: Bipolar disorder prediction on Reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Vasilev, E. (2018). Inferring gender of reddit users. Bachelor thesis, GESIS - Leibniz Institute for the Social Sciences.

XueMing Leng and YiDing Wang (2008). Improving generalization for gender classification. In *2008 15th IEEE International Conference on Image Processing*, pages 1656–1659.