LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

# Bachelor Thesis

Major Computational Linguistics
at Ludwig Maximilians University Munich
The Center for Information and Language Processing

# Cross-genre Author Profiling

submitted by
Ivan Bilan

Supervisor: Dr. Desislava Zhekova
Duration: March 21st – May 30th, 2016

# Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, den 30. Mai, 2016

……………………………………………………………………………………………………….

Ivan Bilan

## Abstract

Author profiling is a task of determining the sociodemographic attributes of an author, including their gender and age. Author profiling may be used for customer-base analysis or in the field of forensics to help identify a criminal. The classification system for gender and age profiling, proposed in this thesis, considers the usage of different parts-of-speech, collocations, connective words and various other stylometric features to differentiate between the writing styles of male and female authors as well as among the different age groups. The proposed approach concentrates on online social media datasets and explores the effectiveness of working with various genres at the same time in comparison to only concentrating on one genre. The system achieves an average 64% and 36% accuracy for gender and age classification respectively on the English datasets in a cross-genre format. Additionally, the same system scores up to 83% accuracy in single genre classification of author's gender on English datasets. Moreover, the system is also suited for Spanish and Dutch text samples. For cross-genre gender classification on the Spanish dataset, the model reaches the performance of 59.38%. A lower result is exhibited for Dutch amounting to only 54.50%.

# Contents

# 1 Introduction

## 1.1 Importance of Author Profiling

With over three billion[1] Internet users and more than a hundred[2] popular social media websites, the number of anonymous users who may abuse the global network is growing. For instance, sexual predators who create fake accounts to stalk people or teenagers illegally using adult social media websites. To identify the age or gender of a user behind the fake account, software solutions that analyze the writing style of the user based on his written texts may be used.

Apart from misuse, social media may be utilized for user base analysis. By analyzing the reviews of company's products or Twitter[3] posts mentioning a certain product, the average age of the consumer may be determined. Furthermore, the analysis may tell whether men or women prefer the product more.

Additionally, finding out the gender and age of the author may also be used for targeted advertising, by placing the ads that would interest the user the most, next to the online text written by them.

Profiling the author of a text may also be useful in the field of forensics, for instance, to analyze a ransom note or a set of anonymous blogs written by a hacker to better understand the sociolect aspect of the suspect.

Author profiling may be used to tackle all these problems, for instance, to help identify a criminal or to analyze the consumer base. The notion of author profiling can be best explained in contrast to the task of author attribution. In the latter, a set of authors and a number of text examples belonging to those authors is given. The aim of author attribution systems is to assign the text samples to the authors who wrote them. An illustrative example of author attribution technique is provided in the FBI Law Enforcement Bulletin by Smith et al. (2002), where a stalker is identified based on the love letters he wrote to the victim in the past.

Author profiling, on the other hand, does not deal with specific authors but with author classes, mainly sociodemographic ones. Gender identification is the most common author profiling task. Age identification is a more complex task and implies attributing the author to a certain age group, for instance, 18 to 24-year-olds or a category of 25 to 35-year-olds. Moreover, with the help of author profiling native language of the author writing texts in a foreign language may be identified, as evident by the research of Koppel et al. (2005). It can also be used to evaluate such personality traits of the writer as emotional stability and neuroticism, whether the author is extrovert or introvert, as explored in PAN15 Shared Task[4], or even predict the level of psychopathy[5] the person writing the texts may demonstrate.

Author profiling analysis may be performed by a trained psychologist, although humans may not catch all the underlying subtle stylistic, linguistic and structural differences between certain classes of writers. In the recent years, machine learning systems have been broadly used for various author profiling problems and have shown promising results on multiple datasets and writing styles. These systems extensively use various stylometric features, features that analyze the linguistic style of the author, as well as different techniques borrowed from information retrieval, data mining, and statistical data analysis.

[1] http://www.internetlivestats.com/internet-users/

[2] https://en.wikipedia.org/wiki/List_of_social_networking_websites

[3] https://twitter.com

[4] http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html

[5] https://www.kaggle.com/c/twitter-psychopathy-prediction

### 1.2 Cross-genre Aspect of Author Profiling

Previous research into author profiling concentrated on building machine learning systems based on a certain genre. In this case, the classification model would be trained and tested on the same genre. Recently a new view on this problem was proposed by the PAN16 Shared Task[6], namely the use of a system capable of applying the learned stylistic and structural differences between author classes of one writing style to a previously unseen one.

This thesis expands the notion of cross-genre author profiling and additionally explores the possibility and effectiveness of a classification system which would work on both single genre and cross-genre datasets.

The existence of an effective cross-genre author profiling system would eliminate the problem of data scarcity since any available dataset in any genre could be used to train the model. Some social media genres are easier to collect and label than the others. For example, users may report their age and gender on a certain blogging website more often than on a site that concentrates on reviews. In this case, collecting an extensive dataset of blogs would take less effort and be more precise than collecting and labeling a dataset of reviews. Furthermore, the collected blog dataset could be used as a training set to build a model capable of profiling the review authors.

### 1.3 Thesis Overview

The main aim of this thesis is to research the effectiveness and a possible implementation of cross-genre author profiling system.

After a short introduction to author profiling in chapter 1, an overview of previous research work in this field is given in chapter 2. Additionally, a description of shared tasks dealing with this classification problem is presented.

Chapter 3 gives an overview of the datasets used to train the model, and the process of attaining the labels for each author is reviewed.

Chapter 4 concentrates on the implementation of the supervised machine learning system used for gender and age author profiling. A detailed overview of techniques and stylometric features utilized for the task is presented.

Furthermore, the model is evaluated on various genres, and the results are analyzed in chapter 6. Ultimately, the last chapter draws the final conclusions regarding the effectiveness of a cross-genre profiling system and approaches possible future improvements of the classification model.

[6] http://pan.webis.de/clef16/pan16-web/author-profiling.html

# 2 Related Work

## 2.1 Early Contributions to Author Profiling

The research conducted by Koppel et al. (2002) may be considered the pioneering work into the area of author profiling. With the help of various stylistic and lexical features, Koppel et al. [1] showed that a gender of the author of an anonymous text may be determined with the aid of machine learning. The classification was run on a small dataset of 566 documents from the British National Corpus (BNC)[7] subsequently achieving 80% accuracy. Moreover, work by Argamon et al. (2003) proves that male and female authors use different writing styles. The research shows that the usage of various parts-of-speech differs considerably throughout male and female writers. Both of these works concentrated on long textual samples belonging to fiction and non-fictional genres of writing.

Research conducted by Schler et al. (2006) explored the effectiveness of automatically profiling the authors of online texts, namely blogs. Moreover, apart from gender classification the work also concentrated on the age of the author and was conducted on an extensive dataset of over 70.000 text samples achieving 80% accuracy for gender and 75% for age classification.

Mukherjee et al. (2010) experimented with various feature combination to find out the gender of the blog authors. The research encompassed a wide array of stylistic and structural features and experimented with an ensemble feature selection algorithm to deploy the most performative feature combination. The task was conducted on some 3.100 blog posts, and the final classification model achieved 88.5% accuracy.

Nonetheless, the early research concerning author profiling concentrated on long text samples. Nowadays more and more attention is given to short social media texts, partly caused by the success of such popular website as Twitter where users are restricted to 140 characters per post.

## 2.2 PAN Shared Tasks

Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) shared task is a workshop at a yearly Conference and Labs of the Evaluation Forum (CLEF)[8] independent event. PAN covers a broad range of text manipulation and classification tasks like author masking, author attribution, author profiling and plagiarism detection.

Author profiling became a part of the shared task in 2013 and concentrated on profiling the authors of blog posts into appropriate gender and age classes. The task included datasets in English and Spanish. Age of the authors was divided into three categories: 13 to 17-year-olds, 23 to 27 and 33 to 47-year-olds. Additionally, PAN13 author profiling task also included a subtask of identifying sexual predators, who have created fake accounts to use them for the purpose of sexual harassment, amongst the authors present in the dataset. The best performance for gender profiling reached 59% accuracy for English and was achieved by Meina et al. (2013), whereas the best result for age classification was obtained by Santosh et al. (2013) and exhibited around 66% accuracy.

PAN14, conducted in 2014, included various other genres. Apart from blogs, hotel reviews and tweets were included. In addition to English and Spanish, Dutch and Italian datasets were a part of the task. According to the task overview by Rangel et al. (2014), the best result for gender classification was achieved on a Twitter dataset with 76% accuracy. Other genres exhibited lower results ranging from 53% accuracy for social media dataset for gender identification to 73%

[7] http://www.natcorp.ox.ac.uk

[8] http://clef2016.clef-initiative.eu

accuracy for the classification of the hotel review authors. The age identification proved to be a more complex task with the highest accuracy of only 35% for hotel reviews in English.

In 2015, PAN15 workshop concentrated on tweets extracted from the popular social media website Twitter[9]. Moreover, the classification task consisted not only of gender and age identification but also was extended to the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The Big Five framework for personality profiling was simultaneously developed by Goldberg (1990) and McCrae et al. (1987) in the late 80s and was based on the early findings in psychology on personality traits by Cattell (1945).

The best performance on the PAN15 shared task for both gender and age identification was achieved by Álvarez-Carmona et al. (2015) with 84% and 79% for gender and age classification respectively for the English language.

PAN16 shared task, conducted in the second quarter of 2016, implies the usage of a dataset of tweets to train a cross-genre author profiling system which can perform successfully on an unseen genre. The classification model described in this thesis is also competing in the PAN16 shared task.

[9] https://twitter.com

# 3 Datasets in Use

## 3.1 Overview of PAN16 Dataset

The primary purpose of PAN16 shared task is to train a classification model on tweet samples in such a way that it will be able to profile the authors of a previously unseen genre. For this task, the classification model needs to be very robust and adapt the observations learned from short and usually stylistically and grammatically malformed tweet samples to work on any possible form of text as well as samples of any length, for example, hotel reviews, blogs or any other writing style. The classifier submission for the PAN16 shared task is expected to be trained on a Twitter dataset gathered by the task chair. Moreover, the task covers more than one language and includes datasets for English, Spanish, and Dutch.

The PAN16 training dataset includes a considerable number of duplicate text samples, most probably due the use of automated web scraping for dataset collection. Table 3.1 gives a detailed overview of the PAN16 shared task's dataset after the duplicate removal. Authors of English and Spanish text samples are labeled with both gender and age, whereas in Dutch, only gender is given. All datasets have an almost equal amount of male and female authors, although the underlying text samples are predominant in the male gender class.

The sample distribution of the age author classes is rather uneven. Most of the authors, as well as the underlying text samples, belong to two age groups, namely 25 to 34-year-olds and 35 to 49-year-olds. The rest of the age classes are considerably underrepresented, as evident, for instance, by the presence of only six authors in the 65-year-olds and over age class in the English dataset in contrast to 181 authors in the 35 to 49-year-olds age group of the same dataset.

## 3.2 Overview of the Custom Dataset

Since the test set for PAN16 is not openly accessible, and its genre is unknown until the end of the shared task's evaluation stage, it can only be speculated which genre will be considered. For this reason, an additional dataset of blogs, hotel reviews, and tweets was put together to test the classifier on various genre combinations. The training stage mainly concentrated on the English language with the intent of the final classification pipeline being adapted to Spanish and Dutch.

**Table 3.1.** PAN16 Training Dataset Breakdown

| **Training Set** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Language** | | **Text Samples** | | | | | **Unique Authors** | | | | |
| **English** | ***Age*** | *18-24* | *25-34* | *35-49* | *50-64* | *65-xx* | *18-24* | *25-34* | *35-49* | *50-64* | *65-xx* |
| | *Samples* | 15725 | 68936 | 79338 | 34668 | 1435 | 28 | 137 | 181 | 80 | 6 |
| | ***Gender*** | *Male* | | *Female* | | | *Male* | | | *Female* | |
| | *Samples* | 111030 | | 89072 | | | 216 | | | 216 | |
| | *Total* | *200102 Text Examples* | | | | | *432 Authors* | | | | |
| **Spanish** | ***Age*** | *18-24* | *25-34* | *35-49* | *50-64* | *65-xx* | *18-24* | *25-34* | *35-49* | *50-64* | *65-xx* |
| | *Samples* | 7146 | 30730 | 66287 | 21449 | 2869 | 16 | 63 | 38 | 20 | 6 |
| | ***Gender*** | *Male* | | *Female* | | | *Male* | | | *Female* | |
| | *Samples* | 70129 | | 58352 | | | 124 | | | 125 | |
| | *Total* | *128481 Text Examples* | | | | | *249 Authors* | | | | |
| **Dutch** | ***Gender*** | *Male* | | *Female* | | | *Male* | | | *Female* | |
| | *Samples* | 33111 | | 33773 | | | 188 | | | 191 | |
| | *Total* | *66884 Text Examples* | | | | | *379 Authors* | | | | |

**Table 3.2.** Custom Training Dataset Breakdown

| Training Set | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Genre** | | **Text Samples** | | | **Unique Authors** | | |
| **Social Media / Blogs** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 20682 | 20860 | 20856 | 1405 | 1400 | 1426 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 31348 | 31050 | | *2117* | *2114* | |
| | *Total* | *62398 Text Examples* | | | *4231 Authors* | | |
| **Hotel Reviews** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 20773 | 20766 | 20781 | 16782 | 14958 | 15948 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 28826 | 33494 | | 26268 | 21420 | |
| | *Total* | *62320 Text Examples* | | | *47688 Authors* | | |
| **Tweets** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 16378 | 16575 | 17655 | 57 | 64 | 66 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 25190 | 25418 | | 104 | 83 | |
| | *Total* | *50608 Text Examples* | | | *187 Authors* | | |

**Table 3.3.** Custom Test Dataset Breakdown

| Test Set | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Genre** | | **Text Samples** | | | **Unique Authors** | | |
| **Social Media / Blogs** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 5160 | 5260 | 5161 | 352 | 352 | 357 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 7694 | 7887 | | *545* | *516* | |
| | *Total* | *15581 Text Examples* | | | *1061 Authors* | | |
| **Hotel Reviews** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 5194 | 5191 | 5196 | 4509 | 4282 | 4354 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 7178 | 8403 | | 7242 | 5903 | |
| | *Total* | *15581 Text Examples* | | | *13145 Authors* | | |
| **Tweets** | ***Age*** | *18-34* | *35-49* | *50-xx* | *18-34* | *35-49* | *50-xx* |
| | *Amount* | 4350 | 3988 | 3676 | 15 | 18 | 17 |
| | ***Gender*** | *Male* | *Female* | | *Male* | *Female* | |
| | *Amount* | 6440 | 5574 | | 23 | 27 | |
| | *Total* | *12014 Text Examples* | | | *50 Authors* | | |

A smaller sample of the “Social Media” (Blogs) dataset from PAN14 Shared Task[10], “Hotel Reviews” from Webis-Tripad-14 dataset[11] and “Twitter” dataset from PAN16 Shared Task are used. Table 3.2 represents the distribution of social media, hotel reviews, and Twitter datasets into age and gender classes, as well as the number of authors and text examples within each training dataset. Table 3.3 shows a similar distribution pattern for the test set.

The dimensionality of the datasets was reduced to balance the gender and age distribution since all of these shared task datasets have uneven age class representation. Moreover, authors from the 18 to 24-year-olds age group have been integrated into the 25 to 34-year-olds class and authors over 65 into the 50 to 64-year-olds class due to a small number of authors and text samples in those categories, which allows for a nearly uniform gender and age distribution.

[10] http://pan.webis.de/clef14/pan14-web/author-profiling.html

[11] https://www.uni-weimar.de/en/media/chairs/webis/corpora/webis-tripad-14/

All in all, over 66 thousand unique authors and nearly 219 thousand text examples are present in the custom dataset. The highest number of authors is present in the hotel reviews and the lowest in tweets.

### 3.3 Dataset Labeling Process

The PAN Shared Task organizers annotated all the datasets manually. Throughout the shared tasks that concentrated on author profiling four datasets have been collected: social media dataset, blogs, hotel reviews, and tweets.

According to the PAN13 Shared Task overview by Rangel et al. (2013), the "Social Media" dataset consists of blogs collected from the Netlog[12] website. The information about gender and age is taken from the user's profile (Rangel et al., 2013).

The blogs dataset, as described in the PAN14 Shared Task overview by Rangel et al. (2014), was collected from LinkedIn[13]. The birth date given on the profile was used to identify the age. In case it was missing, the age was identified with the help of the university degree starting date given in the profile. After clearly distinguishing the age of the authors, the gender is determined by the photo and the name of the user. The annotation was conducted by two independent annotators and the disagreements resolved by the third one (Rangel et al., 2014).

The hotel reviews dataset is derived from the Webis-Tripad-13[14] Dataset. The hotel reviews are collected from the TripAdvisor[15] website, the gender, and age of the review author is given in their user profile (Rangel et al., 2014). For the custom dataset used for cross-genre classification, Webis-Tripad-14[16] dataset was used, which used a similar author classification process for data collection.

The authors in the Twitter dataset were also labeled using the metadata of their profiles, as stated in the PAN14 shared task overview by Rangel et al. (2014). This dataset is acquired from the PAN-RepLab[17] shared task, used for development of the Online Reputation Management systems, and represents the users with high reputation from various domains (e.g. environmental, banking, automotive) (Rangel et al., 2014). Furthermore, according to the overview of the PAN15 shared task, tweet authors have also been profiled through an online test in which they had to state their age and gender (Rangel et al., 2015).

[12] http://www.netlog.com/

[13] https://www.linkedin.com

[14] http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-webis-tripad-13-sentiment/

[15] https://www.tripadvisor.com/

[16] https://www.uni-weimar.de/en/media/chairs/webis/corpora/webis-tripad-14/

[17] http://nlp.uned.es/replab2014/#tasks

# 4 Experimental Setup

## 4.1 General Overview

The input dataset overcomes several levels of transformation before the feature extraction stage, namely preprocessing, lemmatization and stemming, as well as part-of-speech (POS) tagging. Then the most informative words, characters, and POS tags are extracted as frequency vectors with the help of Term Frequency-Inverse Document Frequency (TF-IDF) procedure. Simultaneously, the lemmatized words are distributed into topics with the aid of Latent Dirichlet Allocation. In the succeeding step, custom features are used, and the feature vectors are extracted and scaled. Figure 4.1 is a visual representation of the training process for the cross-genre author profiling classifier. A more detailed description of each step is given in the following subsections.

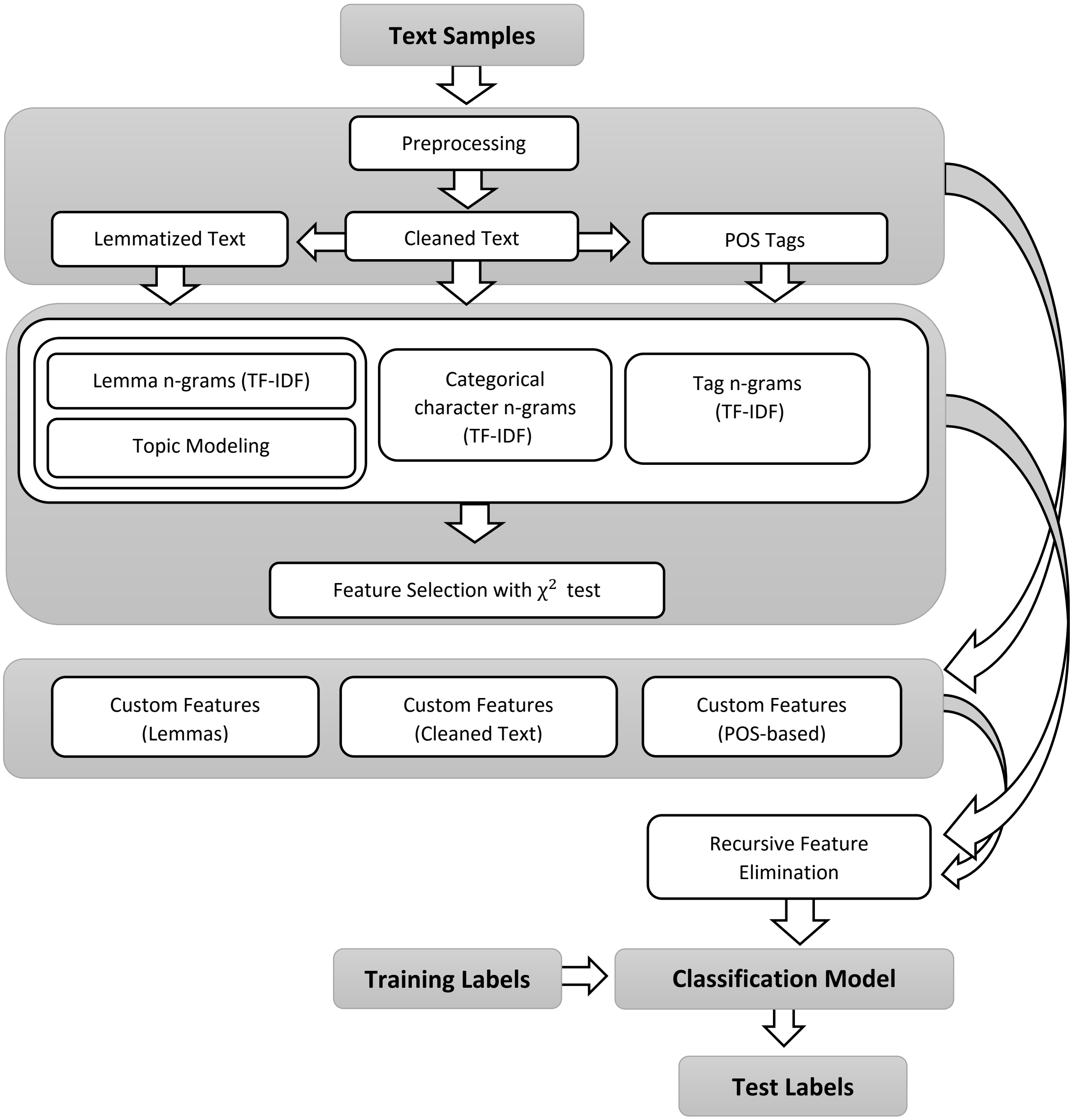

**Fig. 4.1.** Final Classification Pipeline

### 4.2 Preprocessing

Since the datasets have been collected through the use of web scraping, some text samples include partial HTML code and Bulletin Board Code used for formatting. Many authors use web links in their texts, to link to external websites, images and videos. Also, tweet samples include references to usernames of other Twitter users and use hashtags to link their tweets to some common topic.

In the preprocessing step, all forms of web-related code have been removed. All links are normalized to a special token *[URL]*, and all username mentions of a form *@username* are translated to *[USER]*. Hashtags are left unprocessed since they may represent the authors' interests and provide stylistic information. The text was not normalized in any other way to preserve its structure.

### 4.3 Data Transformation

Instead of concentrating on preprocessing the text, various other textual representations of it have been created. Porter Stemmer, a stemming tool first formulated by Porter (1980), namely its NLTK implementation[18], is used to produce a stemmed text. Early experiments showed that the stemming procedure strips off too much morphological and subsequently stylistic information from the words. Lemmatization, on the other hand, allows for a better text representation and provides a more informative vector representation with TF-IDF. TreeTagger[19], introduced by Schmid (1994), is used to lemmatize the text as well as to produce part-of-speech text representation of the input text samples.

The text is also represented as a TF-IDF vector of categorical character n-grams, first proposed by Sapkota et al. (2015) and successfully used for author profiling by Maharjan et al. (2015). This technique analyses the usage of various character groups like prefixes, suffixes, word ending and word beginning character pairs.

### 4.4 Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF), first formulated by Jones (1972), is a procedure widely used in information retrieval and data mining to measure the importance of each word in a corpus. The word's frequency representation is regulated by a weight variable, which is calculated relative to the frequency a particular token appears throughout the whole dataset. Equation 4.1 shows the mathematical representation of TF-IDF presented by Bilenko (2006) and is modified to include additive plus-one smoothing.

$$w_{v_i,\ s} = \frac{N\,(v_i,\ s)}{max\, v_{j\,\in\,s} N\,(v_j,\ s)} \cdot \left(log\left(\frac{N+1}{N\,(v_i)+1}\right) + 1\right) \tag{4.1}$$

In this representation $v_i$ denotes a token in a text sample *s* and $N\,(v_i,\ s)$ a number of times the token appears in *s*, $N\,(v_i,\ s)$ is also referred to as *term frequency*. *N* is the number of all strings in the corpus and $N\,(v_i)$ is the number of strings that include the token $v_i$, or simply put *document frequency*. Additive plus-one smoothing is used to exclude the possibility of multiplication by zero. For the task of cross-genre author profiling a TF-IDF implementation of the scikit-learn machine learning toolkit, introduced by Pedregosa et al. (2011), is used to convert the lemmatized and POS tag text representations, as well as, character n-grams to a matrix of TF-IDF features.

[18] http://www.nltk.org/_modules/nltk/stem/porter.html

[19] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

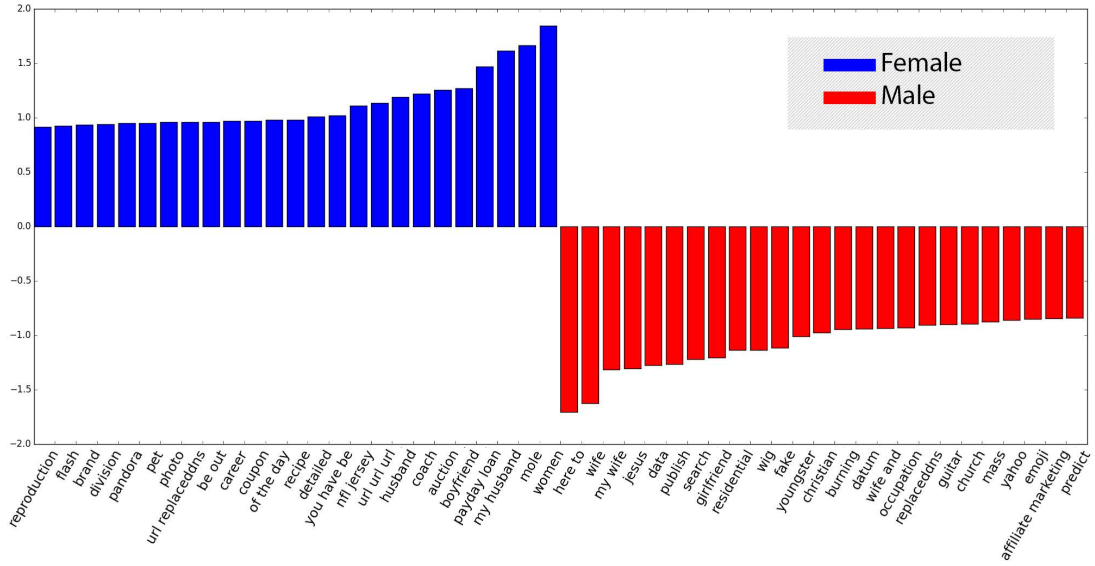

**Fig. 4.2.** Correlation Coefficient of 25 Most Informative Lemma N-grams by Gender Usage

Figure 4.2 shows the correlation coefficient of 25 most informative lemma unigrams, bigrams, and trigrams in the custom English dataset used by women and 25 mostly used by man. It is evident by this representation that both man and women talk a lot about their spouses or that man talk about religion more than women. The representation is based on the custom datasets presented in chapter 3.2 with all genres merged.

According to the initial experiments, TF-IDF vector for the lemma representation shows best results when considering unigrams, bigrams, and trigrams, where the part-of-speech TF-IDF vector additionally considers four-gram POS sequences. The categorical character TF-IDF representation considers only trigram characters.

### 4.5 Topic Modeling

There are several techniques used for topic modeling, among them Latent Dirichlet Allocation (LDA) and word2vec. The latter, first formulated by Mikolov et al. (2013), uses neural networks to put words that have similar meanings into similar clusters. In contrast to word2vec, LDA is a generative statistical model that assigns probability weights to words and according to those probabilities assigns the word to a certain automatically generated topic (Coelho, 2010).

Latent Dirichlet Allocation was first formulated by Blei et al. (2003). For the purpose of author profiling an LDA implementation from gensim vector space modeling and topic modeling toolkit, developed by Řehůřek et al. (2010), is used. LDA process takes the number of possible topics as a mandatory argument. In general, it is a good practice to use more topics for a bigger dataset than for a smaller one. To accommodate to any dataset size, Hierarchical Dirichlet Process (HDP), a modification of LDA, is used to calculate the number of needed topics automatically.

### 4.6 Custom Feature Sets

Apart from TF-IDF vector representation and topic modeling about 40 custom features have been developed for age and gender profiling. Most of these are stylometric features catching the linguistic style of the author. The features are grouped into five clusters: dictionary-based, POS-based, readability features, text structure, and stylistic clusters.

**Table 4.1.** Dictionary-based Features

| Feature Cluster | | Examples per Language | | |
|---|---|---|---|---|
| | **Feature Name** | **English** | **Spanish** | **Dutch** |
| Dictionary-based | Connective Words | *furthermore, firstly, moreover, hence …* | *pues, como, luego, aunque …* | *zoals, mits, toen, zeker …* |
| | Emotion Words | *sad, bored, angry, nervous, upset …* | *espanto, carino, calma, peno …* | *boos, moe, zielig, chagrijnig …* |
| | Contractions | *I'd, let's, I'll, he'd, can't, he'd …* | *al, del, desto, pal', della …* | *m'n, 't, zo'n, a'dam …* |
| | Familial Words | *wife, husband, gf, bf, mom …* | *esposa, esposo, marido, amiga …* | *vriendin, man, vriend, moeder …* |
| | Collocations | *dodgy, telly, awesome, freak, troll …* | *no manches, chido, sale …* | *buffelen, geil, dombo, tjo …* |
| | Abbreviations and Acronyms | *a.m., p.m., Mr., Inc., NASA, asap …* | *art., arch., Avda., Arz., ant. …* | *gesch., geb., nl, notk, mv, vnl …* |
| | Stop Words | *did, we, ours, you, who, these, because …* | *de, en, que, los, del, donde, como …* | *van, dat, die, was, met, voor …* |

**Dictionary-based** feature cluster: The purpose of this set of features is to search the raw text representation for words from a predefined list of tokens. The feature cluster consists of dictionaries of connective words, emotion words, contractions, family related words (a feature proposed by Maharjan et al. (2015)), collocations, abbreviations, and acronyms, as well as stop words. All included dictionaries are available in English, Spanish, and Dutch. A more detailed overview of this cluster with its underlying examples is given in Table 4.1.

**POS-based** feature cluster analyzes the POS text representation, namely the distribution of different parts-of-speech. As evident by Figure 4.3, which represents the usage of conjunctions by male authors throughout the custom training dataset, and Figure 4.4, which shows the representation for the female writers, there is a visible difference in the usage of conjunctions between male and female writers. While the usage distribution per single text is only slightly different, in general, men use conjunctions in their texts more often than women.

Additionally, this set includes a more complex F-Measure feature, first introduced by Heylighen et al. (2002), which can tell how implicit or explicit the text is. The F-Measure is calculated based on the usage of various POS tags in the text. Equation 4.2 gives a more detailed representation of how F-Measure is calculated; all parts of speech in the formula denote the frequency of the POS usage in the text under review.

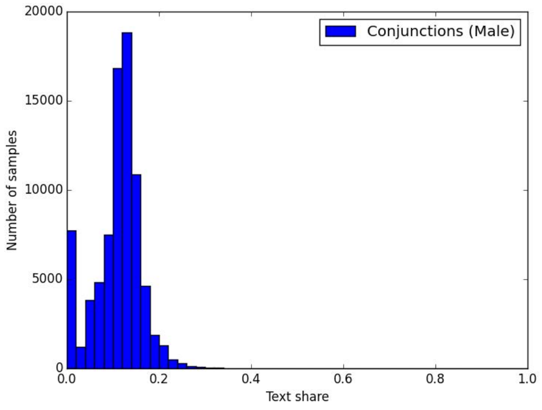

**Fig. 4.3.** Use of Conjunctions by Male Authors

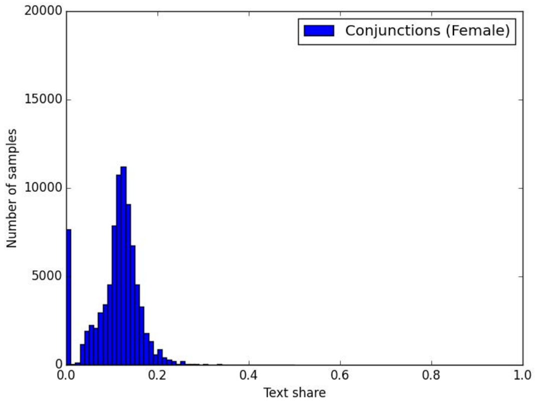

**Fig. 4.4.** Use of Conjunctions by Female Authors

$$F = 0.5 \cdot \Big(\big((nouns + adjectives + prepositions + articles) - (pronouns + verbs + adverbs + interjections)\big) + 100\Big) \tag{4.2}$$

**Readability** features: Various readability features are applied to the raw text representation: Automated Readability Index, SMOG Readability Formula, New Dale-Chall and Flesch Reading Ease Formula. Equation 4.3 represents a formula used to calculate the Flesch Reading Ease, introduced by Flesch (1948). The outcome of the calculation lies between 0 and 100, where values between 0 and 29 classify the text as very confusing to read, values 60 to 69 describe a standard text, and extremes from 90 to 100 label the text as very easy to understand.

$$206.835 - 1.015 \left(\frac{total\,words}{total\,sentences}\right) - 84.6 \left(\frac{total\,syllables}{total\,words}\right) \tag{4.3}$$

The final classification model did not include the readability features since they did not improve the classification results in any way. Various readability indexes are mostly oriented to analyze long text samples and do not perform well on short tweet samples.

**Text structure** feature cluster: Features, belonging to this group, attempt to analyze the structure of the text and consist of type/token ratio, average word length, and the amount of various punctuation marks used in each text sample. Usage of exclamation and question marks are evaluated by separate features, while all other punctuation marks are represented as one feature value. Additionally, the number of linked content tokens and username mentions is considered.

**Stylistic** cluster: This set of features counts the amount of different adjectival and adverbial suffixes in the text samples. First introduced by Corney et al. (2002) for the classification of emails in English, the feature set implies that man use more emotionally intensive adverbs and adjectives, such as *awful*, *dreadfully*, *terribly* (Mukherjee et al., 2010).

### 4.7 Feature Scaling

After all feature vectors are extracted their values are scaled. There are several techniques used for feature scaling. Some of the most widely used are normalization and standardization. Equation 4.4 gives a mathematical formulation of the latter, presented by Raschka (2015), where $x^{(i)}$ is a feature vector sample, $\mu_x$ is sample mean of the feature column, and $\sigma_x$ represents the standard deviation of the feature column.

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{4.4}$$

Equation 4.5, also presented by Raschka (2015), shows the mathematical overview of normalization, with $x^{(i)}$ as a feature vector sample, $x_{min}$ the smallest value in the feature column and $x_{max}$ the largest value in the feature column.

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}} \tag{4.5}$$

The normalization technique scales the feature vectors to a bounded interval although it is not sensitive to the outlier feature values. Through standardization, the feature columns take the form of a normal distribution and the information about outliers is preserved (Raschka, 2015).

In the case of cross-genre author profiling, feature vectors of the training set differ greatly from the feature vectors of the test set, since the lengths of text samples throughout various writing styles range from short tweets to very long samples, as for instance, blog posts. Using normalization or similar form of rescaling would be most suitable for this task, but the information about outlier values would be lost.

To be able to use standardization for this particular task, a form of feature vector pre-scaling is introduced. Pre-scaling implies rescaling the feature columns, which count the number of occurrences of a certain token or stylistic characteristic, relative to the sample length. One solution to this problem is division of the feature column value by the length of the text in tokens. A more comprehensive approach is represented in Eq. 4.6 and implies rescaling of the sample length relative to the smallest mean length of a text sample throughout all possible writing styles that could be represented in both training and test sets, and further division of the feature column value by a rescaled sample length. The rescaled sample length represents the amount of possible smallest sample entities that would fit into the initial text sample. Using this technique the feature column value is always scaled relative to the minimum mean length of all text samples of all represented writing styles. The average length is used instead of minimal or maximum lengths to better represent the sample length distribution of the writing style that has the shortest text samples in the dataset.

$$x_{pre-scaled}^{(i)} = \frac{x^{(i)}}{\left(\frac{len(\varepsilon_i)}{min(\mu_{y_1} \dots \mu_{y_n}) \mid y_n \coloneqq len(\varepsilon_{m_1}) \dots len(\varepsilon_{m_n})}\right)} \tag{4.6}$$

Equation 4.6 gives a mathematical formulation of the feature pre-scaling approach, where $x^{(i)}$ represents the current feature column value, $\varepsilon_i$ is current text sample, $\mu$ stands for mathematical mean, $y_n$ represents a genre and $\varepsilon_{m_n}$ is the text sample of the genre $y_n$. *len()* is a function which, given a text sample, returns its length either in tokens or in characters, which makes this interpretation suitable for both types of features that work on the level of tokens and the ones dealing with character representation.

### 4.8 Feature Selection

As demonstrated by the research of Soler Company et al. (2014), author profiling systems can produce competitive results for gender and age discrimination even with a small subset of features.

Since the use of TF-IDF vector representations for lemma, part-of-speech and character n-grams produces a vector of thousands of features, for the final classification model only a smaller percentile of these features is selected. For this purpose, the chi-square test is used. With its help, the independence of term and class occurrences is tested, and subsequently, the variables that are independent of the class are eliminated. This ensures the selection of only class dependent features and shrinks the number of features allowing the custom features influence the classification increasingly more.

Afterward, the TF-IDF vectors and the custom feature vectors are merged. To further reduce the number of features recursive feature elimination (RFE) may be applied. RFE, first introduced by Guyon et al. (2002) for gene selection in cancer classification, recursively selects smaller and smaller feature sets by assigning weights to each feature and selecting the best combination of features that includes a certain number of maximum features predefined by the user, which brings the best classification result.

RFE showed an increase in performance by around 2% on a considerably smaller subset of the custom dataset with the restriction of 1000 best features. Since RFE requires compute-intensive calculation of all possible feature combinations, it was discarded as it cannot handle large datasets effectively.

### 4.9 Classification

There are various approaches to the author profiling classification. The task implies classifying the author of a given text sample, but in many cases, there is a whole set of documents belonging to one author. This raises the question of how to handle the big number of samples per author. It is possible to concatenate all text samples of each author into one uniform sample, as demonstrated by Șulea et al. (2015). Another approach is to build intra-profile relations between text samples and the author profile, as described by Monroy et al. (2015), or to classify each text sample separately and then classify the author class based on each text sample belonging to the author. The latter approach is used by the classification model described in this thesis.

Moreover, some researchers, including González-Gallardo et al. (2015), suggest that gender and age classification should be considered as a unified task since these two classes are interrelated. Thus, the classification accuracy improves by viewing both of these attributes at the same time and not as a separate classification task. Others, like López-Monroy et al. (2013), consider it a separate task and build different models for gender and age classification.

The presented classification system approaches gender and age classification as separate problems. Although it uses the same set of features for both classification models, the classifier used for gender profiling differs from the one used for age identification.

Gender classification is performed using a form of Support Vector Machine with a linear kernel called LinearSVC[20], which is based on LIBLINEAR classifier presented by Fan et al. (2008). The age identification uses One-vs.-rest classifier based on Logistic Regression[21]. Various other classifiers have been tested for the task, as for instance, Decision trees classifier or Stochastic Gradient Descent, although they all delivered similar results their performance time wise was very slow since a number of dataset samples reaches over 200.000. Both LinearSVC and Logistic Regression perform the needed calculations much more efficiently and still deliver good results.

It is also important to note that throughout the classification process no attempt is made to look for any explicit information about the author in the text samples, for example, a direct mention of the author's age.

[20] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[21] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

# 5 Experimental Results

## 5.1 Experiments on the Custom Dataset

The final PAN shared task submissions are executed and evaluated through TIRA evaluation platform, introduced by Gollub et al. (2012), which returns the results only in the form of accuracy. For this reason, the custom dataset was used to inspect precision, recall and $F_1$–Score.

The custom dataset was used to evaluate various genre combinations, for example, training on Twitter samples and testing the system on blogs or hotel reviews. Additionally, the model was tested on single genre classification using the custom dataset.

Table 5.1 gives an overview of the single genre model performance on the three genres present in the custom dataset. The best result is achieved on the Twitter dataset when classifying the authors by gender, with the result of 74% in terms of $F_1$–Score and a high precision and recall distribution of 80% and 70%. In contrast, the age classification only slightly outperforms the baseline of simply randomly classifying the authors into the three age groups, in this case, 33%, and achieves 34%. These results illustrate that finding stylistic differences between male and female authors is easier than classifying authors into five age groups.

The gender classification results on blogs and hotel reviews are considerably lower than on the Twitter dataset. Although the result of 49% in gender classification on the used blogs dataset, a refined version of the “Social Media” dataset used in PAN13 shared task, is only 10% lower than the PAN13 best model results of 59%. The outcome of the age classification on blogs (49%) and hotel reviews (40%), however, is better than the results on the Twitter dataset and outperforms the baseline by around 8%.

In addition to single genre classification, various genre combinations have been used for training and testing to evaluate the performance in a cross-genre setting. Table 5.2 reviews all cross-genre combinations in detail, taking into consideration precision, recall and $F_1$–Score values. Regardless of the genre combination, the model performs under the baseline for both gender and age classification, achieving an average of 41.6% for gender and 28.3% for age classification in terms of $F_1$–Score throughout all combinations. Such low performance reflects the complexity of the task and suggests that single genre classification is a better solution for author profiling accuracy wise. Nevertheless, such low performance may also indicate that the dataset used for training requires a bigger number of text samples.

Additionally, a mixture of more than one genre was used for training. According to the results given in Table 5.3, the model achieves performance lower than the baseline of simply randomly selection the classes, in this case, 50% for gender and 33% for age classification. In general, the results are considerably lower than when using only one genre for training. Especially noticeable is the fact that the results on the Twitter test set are significantly lower in both cross-genre and multi-genre classifications compared to the single genre setting. The initial result of gender classification in a single genre format on the Twitter dataset of 74% is around 35% better on average throughout all other forms of cross-genre classification. Low performance of the cross-genre combinations reinforces the idea that single genre classification is a much more efficient form of author profiling.

## 5.2. Single Genre Performance on PAN Datasets

In addition to the participation in PAN16 shared task, the model was trained and tested on the PAN14 and PAN15 shared task datasets through the TIRA evaluation system. These tasks concentrated on single genre gender and age classification.

**Table 5.1.** Single Genre Results Measured in $F_1$–Score

| Language | Class | Train Genre(s) | Test Genre | Precision | Recall | $F_1$ | Baseline |
|---|---|---|---|---|---|---|---|
| English | Gender | Twitter | Twitter | 80.00 | 70.00 | 74.00 | 50.00 |
| English | Age | Twitter | Twitter | 35.00 | 36.00 | 34.00 | 33.33 |
| English | Gender | Hotel Reviews | Hotel Reviews | 57.00 | 57.00 | 57.00 | 50.00 |
| English | Age | Hotel Reviews | Hotel Reviews | 40.00 | 40.00 | 40.00 | 33.33 |
| English | Gender | Blogs | Blogs | 51.00 | 51.00 | 49.00 | 50.00 |
| English | Age | Blogs | Blogs | 44.00 | 44.00 | 44.00 | 33.33 |

**Table 5.2.** Cross-genre Results Measured in $F_1$–Score

| Language | Class | Train Genre(s) | Test Genre | Precision | Recall | $F_1$ | Baseline |
|---|---|---|---|---|---|---|---|
| English | Gender | Twitter | Hotel Reviews | 55.00 | 51.00 | 48.00 | 50.00 |
| English | Age | Twitter | Hotel Reviews | 35.00 | 35.00 | 32.00 | 33.33 |
| English | Gender | Twitter | Blogs | 48.00 | 48.00 | 41.00 | 50.00 |
| English | Age | Twitter | Blogs | 49.00 | 36.00 | 25.00 | 33.33 |
| English | Gender | Hotel Reviews | Twitter | 76.00 | 48.00 | 33.00 | 50.00 |
| English | Age | Hotel Reviews | Twitter | 49.00 | 38.00 | 36.00 | 33.33 |
| English | Gender | Hotel Reviews | Blogs | 56.00 | 50.00 | 38.00 | 50.00 |
| English | Age | Hotel Reviews | Blogs | 46.00 | 38.00 | 32.00 | 33.33 |
| English | Gender | Blogs | Hotel Reviews | 51.00 | 49.00 | 48.00 | 50.00 |
| English | Age | Blogs | Hotel Reviews | 35.00 | 35.00 | 31.00 | 33.33 |
| English | Gender | Blogs | Twitter | 76.00 | 56.00 | 42.00 | 50.00 |
| English | Age | Blogs | Twitter | 09.00 | 28.00 | 14.00 | 33.33 |

**Table 5.3.** Multi-genre Results Measured in $F_1$–Score

| Language | Class | Train Genre(s) | Test Genre | Precision | Recall | $F_1$ | Baseline |
|---|---|---|---|---|---|---|---|
| English | Gender | Hotel Reviews, Twitter | Blogs | 52.00 | 49.00 | 39.00 | 50.00 |
| English | Age | Hotel Reviews, Twitter | Blogs | 48.00 | 36.00 | 33.00 | 33.33 |
| English | Gender | Blogs, Twitter | Hotel Reviews | 54.00 | 50.00 | 48.00 | 50.00 |
| English | Age | Blogs, Twitter | Hotel Reviews | 35.00 | 35.00 | 30.00 | 33.33 |
| English | Gender | Blogs, Hotel Reviews | Twitter | 29.00 | 54.00 | 38.00 | 50.00 |
| English | Age | Blogs, Hotel Reviews | Twitter | 43.00 | 32.00 | 18.00 | 33.33 |

**Table 5.4.** Results on PAN15 Datasets Measured in Accuracy

| Language | Class | Train Genre(s) | Test Genre | Model | PAN15 Best | Baseline |
|---|---|---|---|---|---|---|
| English | Gender | Twitter | Twitter | 83.70 | 85.92 | 50.00 |
| English | Age | Twitter | Twitter | 73.53 | 83.80 | 25.00 |
| Spanish | Gender | Twitter | Twitter | 90.99 | 96.59 | 50.00 |
| Spanish | Age | Twitter | Twitter | 66.86 | 79.55 | 25.00 |
| Dutch | Gender | Twitter | Twitter | 79.07 | 96.88 | 50.00 |

Table 5.4 compares the results of the classification system presented in this paper with the best performance for gender and age classification in PAN15 shared task, both the accuracy of the model and the best accuracy of the shared task are averaged accuracies of two validation sets used in the task. In general, the model successfully performed on single genre PAN15 datasets and produced results close to those of the state-of-the-art systems presented in the tasks. For instance, achieving 83.70% for gender profiling on the English dataset, which is only 2.85% lower than the state-of-the-art system presented in the shared task. The classification model also reached much better results on age classification in comparison to other datasets, amounting to 73% accuracy for the English language set. Moreover, the model achieves around 91% accuracy for Spanish and 79% for Dutch in gender profiling. In general, the model displays almost double the performance increase on the single genre PAN15 test sets in comparison to cross-genre combinations of the custom dataset.

A short overview of the evaluation on PAN14 datasets and the model results comparison to the best results of PAN14 is given in Table 5.5, which illustrates the effectiveness of the model on various single genre classifications. The model produces results that are in general lower than the state-of-the-art solutions presented at PAN14, although the discrepancy lies only within one to ten percent for various datasets. The best result of 72.55% achieved on the English hotel reviews dataset in gender classification is only three percent points lower than the best PAN14 performance.

The results of age classification, however, vary considerably depending on the genre of the dataset. In some cases, being 15% lower than the state-of-the-art solution, for instance, when classifying English blogs. On some datasets the model outperforms the best results of PAN14 by a small margin of around half percent points, as illustrated by the age classification on the English Twitter dataset, reaching 51.06% in comparison to 50.65% of the best PAN14 model.

The performance on the Spanish datasets is lower compared to the English classification and the PAN15 state-of-the-art solutions for this language. The best result is exhibited on Spanish Twitter dataset gender classification, reaching 59.11% accuracy, which is around 6% lower than the state-of-the-art solution presented at PAN14. The age identification for Spanish shows 42.86% accuracy for blogs dataset and 47.52% for the Twitter dataset.

To further validate the results on single genre datasets, the model was tested on a blog dataset collected by Mukherjee et al. (2010). The dataset consists of 3.100 blogs with each author represented by one blog sample. The system proposed by Mukherjee et al. (2010) achieves 88.56% accuracy. The dataset was divided into training and test set with 2573 samples and 644 samples respectively in each set. The results of the classification are given in Table 5.6. In general, the model achieves 71.89% in terms of accuracy which is 16.67% lower than the solution by Mukherjee et al. (2010). However, the discrepancy may be explained by a small number of authors in the dataset. Nevertheless, the model outperforms the baseline achieved by random class selection, in this case, 50%, by 22%.

**Table 5.5.** Results on PAN14 Datasets Measured in Accuracy

| Language | Class | Train Genre(s) | Test Genre | Model | PAN14 Best | Baseline |
|---|---|---|---|---|---|---|
| English | Gender | Blogs | Blogs | 62.05 | 67.95 | 57,69 |
| English | Age | Blogs | Blogs | 30.45 | 46.15 | 16.00 |
| English | Gender | Twitter | Twitter | 61.86 | 73.38 | 59.74 |
| English | Age | Twitter | Twitter | 51.06 | 50.65 | 27.92 |
| English | Gender | Hotel Reviews | Hotel Reviews | 72.55 | 75.58 | 66.26 |
| English | Age | Hotel Reviews | Hotel Reviews | 35.99 | 35.02 | 27.53 |
| Spanish | Gender | Blogs | Blogs | 42.86 | 58.93 | 53.57 |
| Spanish | Age | Blogs | Blogs | 40.18 | 48.21 | 16.07 |
| Spanish | Gender | Twitter | Twitter | 59.11 | 65.56 | 47.78 |
| Spanish | Age | Twitter | Twitter | 47.52 | 61.11 | 46.67 |

**Table 5.6.** Results on Blog Dataset (by Mukherjee et al.) in Accuracy

| Language | Class | Train Genre(s) | Test Genre | Model | Mukherjee et al. | Baseline |
|---|---|---|---|---|---|---|
| English | Gender | Blogs | Blogs | 71.89 | 88.56 | 50.00 |

The dataset includes gender labels only. Thus, the classification model does not evaluate the age distribution.

### 5.3 Final PAN16 Shared Task Results

In PAN16 shared task, the model is trained on a training set of Twitter samples. Although the detailed sample distribution of the test set and its genre are unknown, it is certain that the test set represents a different genre dataset than Twitter. The test set can only be accessed indirectly through TIRA evaluation system which returns the results of the final assessment on the test set. Additionally, the results of all evaluations are returned as accuracy, which does not allow for inspection of precision and recall values. The test set consists of two validation sets and most probably each validation set represents a different genre. Table 5.7 provides a detailed overview of the final results achieved on both validation sets. The final classification model produces some promising results despite the complexity of the task achieving 64% average accuracy throughout all English PAN16 datasets for gender classification. The performance on test set 1 exhibits 53.74% while it reaches 74% accuracy on test set 2.

The results for age classification reached only 36.95% on average for English which reflects the complexity of author profiling into five age groups. Test set 1 and test set 2 exhibited 29.02% and 44.87% accuracy respectively.

A performance discrepancy of around 20% between each validation set leads to an assumption that each test set represents a different genre. The results on the first validation set are worse than the results on the second one, which may indicate that the first set represents a dataset similar to the “Social Media” dataset since throughout all previous PAN shared tasks similar low results have been exhibited on this dataset. In this case, the second validation set probably consists of either blog posts or hotel reviews.

**Table 5.7.** Final PAN16 Results Measured in Accuracy

| Language | Class | Train Genre(s) | Test Genre | Test Set 1 | Test Set 2 | Average | Baseline |
|---|---|---|---|---|---|---|---|
| English | Gender | Twitter | Unknown | 53.74 | 74.36 | 64.05 | 50.00 |
| English | Age | Twitter | Unknown | 29.02 | 44.87 | 36.95 | 16.00 |
| Spanish | Gender | Twitter | Unknown | 56.25 | 62.50 | 59.38 | 50.00 |
| Spanish | Age | Twitter | Unknown | 23.44 | 46.43 | 34.94 | 16.00 |
| Dutch | Gender | Twitter | Unknown | 55.00 | 54.00 | 54.50 | 50.00 |

Results on Spanish and Dutch are much lower in comparison to English. The model reaches 59.38% for gender classification in Spanish and 34.94% on age. Nevertheless, the model still outperforms the baseline of randomly selecting author classes, in this case, 50% for gender and 16% for age classification.

The results on Dutch are similar for both datasets and exhibit 54.50% average accuracy for gender classification. The Dutch dataset does not include age labels.

Lower results on Spanish and Dutch in comparison to the English model performance may be explained by the fact that during the training phase the main focus lied on the English language oriented features. The features included in the model do not deal with any specific language traits of either Spanish or Dutch. Only certain features used for English are adapted for Spanish and Dutch: translated dictionary-based features and adjectival suffixes present in the language. All other feature clusters are similar for all three languages.

To this time, there are no results with which the current model can be compared since the task of cross-genre author profiling has not been performed by other researchers until now. Other teams are participating in PAN16 shared task, but the task overview will only be available during the Conference and Labs of the Evaluation Forum (CLEF) event in September 2016.

# 6 Conclusion

## 6.1 Future Work

The system may be further improved in various ways. Firstly, more attention needs to be paid to Spanish and Dutch since the included features are only tailored to English and simply adjusted to work on other languages represented in the shared task. Features that consider language specific linguistic markers may considerably improve the general performance of the system for the given language.

Secondly, some form of text sample-author profile interrelation can be tried out since the current model considers each text sample as a separate entity which does not correlate with other text samples belonging to the author.

Additionally, the system may be improved by refining the feature sets. Current dictionary-based features may perform better with a larger set of tokens. For instance, the collocations list for the English language currently includes only seventy tokens and should be enriched. All dictionary-based features for Spanish and Dutch should be reviewed by a native speaker of each language for more credibility.

Another improvement in accuracy could be achieved by using a meta-classifier, a classifier that based on the classification output of stacked classifiers working on the same classification problem chooses the most suitable class for each sample depending on the produced confidence score.

## 6.2 Final System Analysis

The task of cross-genre author profiling is a more complex one in comparison to the single genre profiling. The classification model needs to generalize learned observations throughout various text genres.

The final classification model showed promising results on a cross-genre aspect of author profiling regardless of the complexity of the task. The model reached 64% accuracy for gender identification in cross-genre profiling of English authors and an average of 36.9% for age classification. The results on Spanish (achieving 59.38% for gender classification) and Dutch (54.50% for gender profiling) datasets are lower since the system was mainly tailored for the English text samples.

Ultimately, the model reached 64.05% accuracy for gender identification in PAN16 cross-genre classification and 83.70% accuracy for single genre profiling in PAN15 on English datasets, which is indicative of the fact that the same model may perform well on both cross-genre and single genre datasets. The cross-genre aspect of the classification model needs to be considerably improved to compete with the single genre performance since the discrepancy between the two aspects of classification is around 20% on average for gender profiling. Moreover, the age classification on cross-genre datasets presents a more complex problem than on single genre profiling since the results achieved on PAN16 cross-genre English datasets are around 40% lower than the results reached on PAN15 single genre English datasets.

In general, the model achieved better results on PAN15 and PAN16 shared tasks than on the custom dataset. Such outcome may indicate that the presented classification system works best on a larger dataset since the custom dataset includes considerably lower amount of text samples than most of the training datasets presented in the PAN shared tasks.

Additionally, the results on cross-genre classification were higher when using only one genre for training as opposed to merging multiple genres in the training set. This indicates that the considerable differences between the genres negatively affect the effectiveness of the model when using more than one genre for training.

All in all, the research into the effectiveness of cross-genre author profiling carried out in this thesis shows that single genre author profiling is still a more efficient way of author classification. Customizing a model for each genre works best since the writing style, structure and vocabulary in use may be better captured by the classifier using genre-specific features. In the case of cross-genre classification, the model requires the use of more general features that apply to multiple genres. As evident by the presented results, single genre gender classification performance is on average around 20% better than cross-genre author profiling results throughout all shared task datasets. The cross-genre performance on the custom dataset is even lower than the performance on PAN16 datasets, and since the custom dataset includes fewer samples than the shared task datasets, such performance also indicates that when using more text samples, the cross-genre performance tends to improve.

Nevertheless, cross-genre classification still outperforms the baseline on PAN16 datasets and shows potential for further research in this field. The development of a uniform model for both cross-genre and single genre classification is a much more time-efficient solution than developing and maintaining multiple systems for each genre.

## References

Álvarez-Carmona, M., López-Monroy, P., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Escalante, H. (2015). INAOE's participation at PAN'15: Author Profiling task - Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers.* Toulouse, France.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse, 23*(3), 321-346.

Bilenko, M. (2006). Learnable Similarity Functions and Their Application to Record Linkage and Clustering. *Citeseer, 2003*(3), 449-467.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3*(1), 993-1022.

Cattell, R. B. (1945). The description of personality: Principles and findings in a factor analysis. *The American Journal of Psychology*, 69-90.

Coelho, L., & Richert, W. (2015). *Building Machine Learning Systems with Python, Second Edition.* Packt Publishing Ltd.

Corney, M., De Vel, O., Anderson, A., & Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. *Proceedings - Annual Computer Security Applications Conference, ACSAC, 2002-January*, pp. 282-289.

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning, 9*, 1871-1874.

Flesch, F. (1948). A new readability yardstick. *The Journal of applied psychology, 32*(3), 221-233.

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology, 59*(6), 1216-1229.

Gollub, T., Stein, B., Burrows, S., & Hoppe, D. (2012). TIRA: Configuring, executing, and disseminating information retrieval experiments. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, 151-155.

González-Gallardo, C., Montes, A., Sierra, J., Núñez-Juárez, J., Salinas-López, A., & Ek, J. (2015). Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams - Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers*.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1-3), 389-422.

Heylighen, F., & Dewaele, J. (2002). Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science, 7*(3), 293-340.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing, 17*(4), 401-412.

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an Author's Native Language by Mining a Text for Errors. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, (pp. 624-628). Chicago, Illinois, USA.

Lim, W., Goh, J., & Thing, V. (2013). Content-centric age and gender profiling - Notebook for PAN at CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* Valencia, Spain.

López-Monroy, P., Montes-y-Gómez, M., Jair Escalante, H., & Villaseñor-Pineda, L. (2014). Using Intra-Profile Information for Author Profiling - Notebook for PAN at CLEF 2014. *CLEF 2014 Evaluation Labs and Workshop - Working Notes Papers.* Sheffield, UK.

López-Monroy, P., Montes-y-Gómez, M., Jair Escalante, H., Villaseñor-Pineda, L., & Villatoro-Tello, E. (2013). INAOE's participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* Valencia, Spain.

Maharjan, S., & Solorio, T. (2015). Using Wide Range of Features for Author Profiling - Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers.* Toulouse, France.

Meina, M., Brodzińska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based Classification for Author Profiling Using Various Features - Notebook for PAN at CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* Valencia, Spain.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, (pp. 1-12).

Mukherjee, A., & Liu, B. (2010). Improving gender classification of blog authors. *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (pp. 207-217).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology, 54*, 547-577.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130 -137.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers.* Toulouse, France.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W. (2014). Overview of the 2nd Author Profiling Task at PAN 2014. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers.* Sheffield, UK.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the Author Profiling Task at PAN 2013. *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* Valencia, Spain.

Raschka, S. (2015). *Python Machine Learning.* Packt Publishing Ltd.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author Profiling: Predicting Age and Gender from Blogs - Notebook for PAN at CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.* Valencia, Spain.

Sapkota, U., Bethard, S., y Gómez, M., & Solorio, T. (2015). Not all character n-grams are created equal: A study in authorship attribution. *2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, (pp. 93-102).

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, (pp. 199-205).

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, (pp. 44-49). Manchester, UK.

Smith, S., & Shuy, R. (2002, April). Forensic Psycholinguistics: Using Language Analysis for Identifying and Assessing Offenders. *FBI Law Enforcement Bulletin, 71*(4), 16-21.

Soler Company, J., & Wanner, L. (2010). How to Use Less Features and Reach Better Performance in Author Gender Identification. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, (pp. 1315-1319).

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Retrieval. *Journal of Documentation, 28*(1), 11-21.

Șulea, O.-M., & Dichiu, D. (2015). Automatic Profiling of Twitter Users Based on Their Tweets - Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers*.

## List of Figures

## List of Tables