INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

Cross Genre Author Profilling using Syntactic N-Grams

T E S I S

Que para obtener el grado de:

Maestría en Ciencias de la Computación

P R E S E N T A:

Iqra Ameer

Directores de Tesis:

Dr. Grigori Sidorov

Dr. Rao Muhammad Adeel Nawaab

Junio 2017

SIP-14 bis

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 01 del mes de junio de 2017 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Cross Genre Author Profilling Using Syntactic N-grams"

Presentada por el alumno:

AMEER		IQRA
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	1	5	1	4	1	6

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA
Directores de Tesis

Dr. Grigori Sidorov

Dr. Rao Muhammad Adeel Nawab

Dr. Alexander Gelbukh

Dr. Ildar Batyrshin

Dr. Olga Kolesnikova

Dr. Marco Antonio Moreno Ibarra

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Marco Antonio Ramírez Salinas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de Mexico el día 12 del mes Junio del año 2017, el (la) que suscribe Iqra Ameer alumno (a) del Programa de Maestría en Ciencias de la Computación con número de registro B151416, adscrito a Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Grigori Sidorov y cede los derechos del trabajo intitulado Cross Genre Author Profilling using Syntactic N-Grams, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección iqraameer133@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Iqra Ameer

Nombre y firma

DECLARATION

I hereby declare that I have implemented this thesis completely on the basis of my personal efforts under the guidance and supervision of Dr. Grigori Sidorov and Dr. Rao Muhammad Adeel Nawab. All the sources used in this thesis have been cited and the contents of this thesis are not plagiarized. No portion of the work presented in this thesis has been submitted in support of any application for any other degree of qualification to this or any other university or institute of learning.

_____

Iqra Ameer

Summary

The process of automatic identification of an author's demographic traits like gender, age, native language, geographical location, personality type and others from his/her written text is termed as author profiling (AP). Currently, it has engaged the research community due to it's promising applications in security, marketing, forensic, fake profiles identification on online communal networks. A variety of benchmark corpora (English text) released by PAN shared task is used to perform our experiments. This study presents a Content-based approach for detection of author traits (age group and gender) for same-genre and for cross genre author profiles as well. Cross genre- the nature of the training genre is different than testing genre.

In our proposed approach, we used different set of features including syntactic n-grams of part-of-speech tags, traditional n-grams of part-of-speech tags, combination of word n-grams and combination of character n-grams. We tried a range of classifier for several profile sizes. We used as baseline word uni-grams and character tri-grams. Our results surpassed the state-of-the-art (same corpora) and baseline methods as well when applying combination of word n-grams for both age group (0.496) and gender (0.734) classification.

Abstract

The process of automatic identification of an author's demographic traits like gender, age, native language, geographical location, personality type and others from his/her written text is termed as author profiling. We are living in the era where technology is growing rapidly and arising many challenging problems for researchers one of the such problems is author profiling. The problem of author profiling has become an important problem in the fields like linguistic forensics, marketing and security. Now most of the text is online. People write and share their opinions and ideas behind the curtain of anonymity. In recent years, online social setups like Twitter, Facebook, Blogs, Hotels Review etc have extended remarkably and have allowed lots of users of all age groups to develop and support personal and professional relations.

However, a shared characteristic of these digital bodies is that it is easy to provide a wrong name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new options to prepare their victims. So, the aim of this research is to predict the demographic traits of the authors for a benchmark existing corpus based on Twitter, Hotel Reviews, Social Media and Blogs' profiles. We explored state of the art techniques for detecting three author traits including age and gender. We used four set of features including Syntactic n-grams of part-of-speech tags, Traditional n-grams of part-of-speech tags, Combinations of word n-grams, Combinations of character n-grams. To detect an author's demographic information from his content we applied information gain as feature selection method to select most discriminated set of features. We used word uni-gram and character three-gram as baseline approach. We compared our results with baseline and state-of-the-art results on the same corpora as well. Evaluation was carried out using accuracy measure. Results showed that these approaches are useful in detecting different author traits and performance improves when Combination of word n-grams used.

Acknowledgment

All praises are for Allah Almighty, the Most Gracious and Most Merciful, who gave me strength to complete this task. Nothing could have been possible without His blessings.

I would like to pay special thanks to my supervisor **Dr. Grigori Sidorov** for his continuous technical and intellectual support, guidance and most important his precious time. His cooperation leads me to this success. I would like to thank my co-supervisor **Dr. Rao Muhammad Adeel Nawab** whose supervision, support and guidance helped me a lot in completing the given task. I would like to appreciate **Dr. Grigori Sidorov**, **Dr. Alexander Gelbukh**, **Dr. Ildar Batyrshin, Dr. Marco Antonio Moreno Ibarra** and **Dr. Olga Kolesnikova** for serving on my committee. I must also thank to one of my friends **Ifrah Pervaz** listening, offering me advice, and supporting me through this entire process.

I would like to admit that I owe all my achievements to my truly, sincere and most loving parents, sisters and friends who mean the most to me, and whose prayers have always been a source of determination for me. They have always supported and encouraged me to do my best in all matters of life.

Iqra Ameer

Contents

Nomenclature

AP	Author Profiling
Char-gram	Character n-gram
CV	cross validation
HR	Hotel Reviews
IG	Information Gain
ML	Machine Learning
NB	Naive Bayes
POST	Part-of-Speech Tags
RF	Random Forest
SM	Social Media
sn	Syntactic n-grams
StArt	State-of-Art
tn	Traditional n-grams
Wn-gram	Word n-gram

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Author Profiling

Author profiling (AP) is the identification process of a person's gender, age, native language, personality traits and other demographic information from his/her written text [21]. We are living in the era where technology is growing rapidly and arising many challenging problems for researchers one of the such problems is author profiling. Now most of the text is online. People write and share their opinions and ideas behind the curtain of anonymity. The problem of AP has become an important problem in the fields like linguistic forensics, marketing and security.

Authorship analysis can be of two types:

- **Author verification tasks** where the style of individual authors is examined, to check whether a writing belongs to specific author or not.

- **Author profiling** distinguishes between classes of authors studying their socialist aspects, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, education, profession or personality type. So AP can simply be defined as given the set of texts you have to identify age group, gender, profession, education, native language and similar personality traits.

1.2 Importance and Applications of Author Profiling

In recent years, online social setups like Twitter, Facebook, Blogs, Hotels Review etc have extended remarkably and have allowed lots of users of all age groups to develop and support personal and professional relations. However, a shared characteristic of these digital bodies is that it is easy to provide a wrong name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new options to prepare their victims. When trying to detect these internet predators, both law enforcement agencies and social network moderators are faced with two main problems: *(i)* the huge number of profiles and communications on social setups make manual analyses almost impossible and (*ii*) internet predators frequently create a false identity, posing as youths, in order to make interaction with their victims. Therefore, proficient automated systems for identity uncovering and inspection are becoming essential.

1.3 Motivation

In everyday life, Due to the rapid reproduction of social media technologies (twitter, facebook, blogs etc), it has been possible for human being to create online co-unities to share information, ideas, personal messages etc. They are accessible to commit the crimes like hide the original identity, wrong information, identification concealing etc.

The main encouragement behind our research is the use of Part of Speech Tag based Syntactic n-grams and Part of Speech Tag based Traditional n-grams to identify the true age and gender of the author of given a document.

1.4 Thesis Focus

The main aim of this thesis is to explore that how Traditional n-grams (tn-grams) of POST and sn-grams of POST as feature helps in conveying profile of a writer. The concept of syntactic n-grams is illustrated in [51, 52, 50]. We will see how sn-grams of POS tags are different from traditional n-grams in the manner of what elements are considered neighbors.

Our technique will build an image of an author's style by using the information enclosed in dependency trees for sn-grams . This information will characterize as syntactic n-grams of POST and will use to conform a vector space. We will also see how traditional n-grams of POST can be helpful in AP task. We will use the supervised machine learning approach. We will describe the features that will use and the engaged supervised machine learning algorithm.

Moreover in our project we also deal with different machine learning techniques and methods for training the modal and compare the results to identify the best and most suitable techniques for our research work. So we can utilize those machine learning techniques to distinguish author's age and gender. Our focus is to analyze as much textual data as we can and classify techniques to identify writer's age and gender.

As we have mentioned above AP is a vast field covering different aspects related to personality, behaviors and emotions of author, our main focus is covering mainly two aspects i.e. age and gender of AP for *PAN 2014* and *PAN 2016* corpora. We will only deal with English Language corpora.

1.5 Problem Statement

The purpose of this research is to see how an author model sentences at syntactic level, so in this way syntactic n-grams of part of speech tags can conquer the topic dependency that traditional n-grams go through. The syntactic n-grams were used in other related tasks such as Author Verification [42], automatic English as second language grammar correction [51], but the main contribution of this thesis is analysis of sn-grams of POST[1] and tn-grams of POST [2]. It will be interesting to see how they will behave in Author Profiling.

1.6 Scope of the Study

Author profiling task has a rising significance in research field belonging to the scientific community. It has immense applications in various fields like marketing,

[1] the elements of syntactic n-grams are POS tags
[2] the elements of traditional n-gram are POS tags

intelligence, forensics, security related to defense. The aim of this study is to analyze and predict the different demographic traits i.e age and gender of author from the Pan-14, Pan-16 corpora which consists sub-corpora of Blog's posts, Hotel Reviews, Social Media and Twitter Tweets. We will extract two types of features i.e syntactic n-grams of part-of-speech tags and traditional n-grams of part-of-speech tags. We will apply content based feature selection method, automatic classification techniques and will evaluate the performance of different classifiers and feature selection methods on mentioned corpora.

1.7 General Objective

General objective of the thesis is: Develop a method for automatic detection of author traits related to the age and gender.

1.8 Specific Objectives

We want to consider how the use of words in our daily language on social networking reflects our personality, thoughts and behaviors.

This thesis aims to achieve the following specific objectives:

- Explore the problem of same genre and cross genre for author profiling.
- Collection of the benchmark corpora from different genres (Social Media, Hotel Reviews, Twitter and Blogs) for cross genre prediction of authors's age and gender.
- Preprocessing on collected corpora in order to make the text more meaningful after removing tags e.g. HTML tags, URLs.
- Features extraction for preprocessed corpora. We planned to extract Syntactic and Traditional n-grams of Part-of-Speech tags, Word and character n-grams.
- Design the experiments for preprocessed corpora.

- Perform the experiments on benchmark corpora to predict author's age and gender.

- Evaluation on four benchmark corpora from different genres e.g. Social Media, Hotel Reviews, Twitter and Blogs.

1.9 Expected Contributions

The expected scientific contributions of the thesis are :

1.9.1 Expected Scientific Contributions

We planned to achieve following scientific contributions:

- Comparison of different machine learning algorithms for author's age and gender for various corpora and for cross genre conditions .

- Comparison of various feature sets on a range of benchmark author profiling corpora on different genres including Social Media, Hotel Reviews, Twitter and Blogs.

- Comparison of results for writer's age and gender with baseline and state-of-the-art results on same corpora.

1.10 Thesis Outline

Rest of the thesis is organized as follows.

Chapter 2 providing an overview of the existing work in Author Profiling and also explains the existing bench mark AP corpora. It is not only provides background information, the possible methods to adopt in, the current study, highlights the research gap along with explaining available techniques for Author Profiling. This chapter will also represents the table of results of literature review study.

Chapter 3 This chapter describes state-of-the-art approaches for Author Profiling. Following that an overview of their used set of features and summery of their achieved results.

Chapter 4 will Present the proposed approach for AP as a multi-label classification problem. Also explain creation of Syntactic n-grams of Part of Speech Tags and Traditional n-grams of Part of Speech Tags. This chapter also aims to demonstrate the experimental setup how our proposed technique can be used for the development and analysis of the author profile detection systems

Chapter 5 . It will explain and analyze the results of our experiments to predict the age and gender of the authors for mentioned corpora.

Chapter 6 will conclude the thesis, final contributions and explain the possible future work.

Chapter 2

Literature Review

To do an exhaustive research on a topic, we should be familiar with related existing work. We should know till which extent the topic has already been explored. Only then, we will be familiar with the existing systems and features they are lacking. The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as linguistics, psychology and computational linguistics. Pennebaker [39] analyzed how a writing style is relate with a person's attributes such as age and gender.

2.1 Introduction

In this chapter we discussed existing techniques for Author Profiling, benchmark corpora which can be used to evaluate the performance of AP and evaluation measure used for AP task.

2.2 Authorship Problem

In old times electronic media was not that much sophisticated, so there was no requirement for its analysis. Now with the advancement in web technology the utilization of electronic media is growing with the passage of time. The usage of Internet media like emails, blogs, websites and research articles has increased that's way the amount of text is incremented so now the need is to identify "who is who" has also become important and it's a field of growing interest these days.

Authorship attribution is a extensive term which is further divided into (i) Author

Profiling (*ii*) Author Identification (*iii*) Author verification and *(iv)* Plagiarism Detection. Author profiling has been explained in Section 1.1 and age and gender identification are sub tasks of AP and they are used to narrow down the suspects. They are also used for classification of texts into classes. Since the focus of this study is on age and gender identification for English language, this literature review will discuss existing corpora, techniques and evaluation measures for age and gender identification.

2.3 Existing Methods for Author Profiling

To our best knowledge the three most popular author profiling techniques are used by the researchers [47].

- Topic based approach
- Stylistic based approach
- Content based approach

These techniques are widely used by researcher for authorship attribution problem. The following sections presents an overview of the Topic, Stylistic and Content based techniques.

2.3.1 Topic Based Approach

This technique is used to find out frequently appearing words in a document from which we identify that document is talking about a certain topic [43]. For example you have a document which contains an article about smart phones. In this document words like Samsung, IPhone, Nokia, BlackBerry, Windows Phones, RAM and charger etc. will appear most frequently which reflects about the topic of the document. Similarly a document written by a female author might have words like cosmetics, clothes and parties etc. whereas male may have discussion about football, politics and technology etc. A document contains different topics and the ratio of these topics varies. For example, a document that describe 5% about female and 10% about male choices so there would probably be 90% male related terms than female. A topic model follows the same idea for extracting

hidden patterns from large collections of documents. Poulston *et al.* [43] implemented topic models by using (LDA) and n grams for determining different profiles of authors. Marquardt *et al.* [28] applied "MRC" and "LIWC" features to separate occurrences of each word related to various psycholinguistic ideas e.g. motion, imagery, concreteness, emotion familiarity, religion and many others.

Topic based approach is statistical approach where statistics of words occurring in documents predict the topic of the document. Frequency of psycholinguistic related words was calculated using LIWC and MRC features [59].

2.3.2 Stylistic Based Approach

In stylistic text analysis writing style of the writer is analyzed and it is used by many researchers in the past for the task of authorship attribution problem. Stylistic text analysis include different stylistic features like frequencies, punctuation, HTML, readability measure, parts-of-speech, and different statistics [13]. It also contains "function words" (words that are content independent) [32], syntactic features [5] or complexity based 9 features such as word and sentence length [57]. Specifically, the function words with parts-of-speech when get combined has proved to be quite successful [25, 3].

2.3.3 Content Based Approach

The other most popular approach for text analysis is content based methods. Content of the text provide some useful information which can be used to identify both age and gender of the author. Schler *et al.* [49] gathered corpora of over 71,000 blogs and fetched content based features, discussed how content of the blog post reflects personality traits of the author. Experiments preformed to predict the age and gender of the blog author.

Pennebaker *et al.* [39] applied content-based techniques for age and gender detection. Male and female are opposite genders and this difference also reflect in their writings. Male authors are more interested in games, politics, news while female authors love to talk about cooking, shopping and parties. For instance, a text that contains content related to squash is more likely to be written by a male author rather than a female. In a text occurrence of words like Macbook, BMW,

Football etc. Thus the occurrence of words like these will increase the chances of it being written by male rather than female. Similarly, occurrence of words or phrases like bridal shower, mother in law, shoes promotions etc will increase the chances of it being written by female. So content based features can be useful to differentiate between male and female authors [49]. Moreover, females use more adjectives and adverbs while writing as compared to males [39]. It has been noticed that youngsters are more interested to talk about video games, school, those who lies in 20s like to have gossip about college life, movies and individuals of 30s mostly tend to write about their occupation, marriage life and politics. Along with these lines, content based components are imperative to recognize writings having a place with various age groups. Content based methodology incorporate components like bag of words, words n-grams, term vectors, named substances, lexicon words, slang words, constrictions and conclusion words [47, 46]. Content based techniques are evaluated using Chi Square, Information Gain, and Gain Ratio tests.

2.4 Existing Corpora for Author Profiling

This section explains the corpora in detail used for author profiling task.

2.4.1 PAN 2013 Author Profiling Corpus

PAN-13 corpus was introduced in 2013 with thousands of blog posts. Themes of these blog posts allow us to analyses standard styles of different authors. The quality and different style of these texts made the authorship task conclusive for determining the age and gender of the anonymous profiles of the authors. For example, women most of the time talk about dresses, makeup or jewelery and men tend to talk about the sports, cars and politics.

Blog messages are used daily for search engine optimization and can be automatically generated by robots or announcement chat bots. These blogs can be used on social media for open discussion about sexuality and some can also break the line and use these systems behave badly and novice conversations that may result in sexual harassment. Along with this, there are many reasons make it important to uncover the fake profiles. Therefore, *PAN* decided to check the validity of the

author by using author profiling approaches which includes identification of gender predator. *PAN* considered online open and public repositories such as Netlog with posts marked with blogger area such as age and gender. Once identified, *PAN* grouped posts by its bloggers, and these bloggers are linked at least with more than 1,000 words of their posts. The posts selected for each blogger, so that the realistic assessment can be drawn from the framework. They distributed the collection into the following portions: education, training, early evaluation and final test. The Authors were carefully splitted in portions by putting each author in exactly in one portion at least. For categorization of different ages of authors *PAN* followed what has already done and included three categories i.e. 10s (13-17), 20s (23-27) and 30s (33-47) [47].

Table 2.1: Statistics of PAN 2013 Corpus

Traits		**Genre**
		Hotels Reviews
Age	10s	17200
	20s	85799
	30s	133597
Gender	Male	118296
	Female	118300

2.4.2 PAN 2014 Author Profiling Corpus

For the purpose of analyzing different author profiling techniques and methodologies, they have created the corpus for four categories i.e. Twitter, blogs, social media and hotel reviews. The corpus is arranged in XML files, one per author. Each author was marked with age and gender information. In *PAN-AP-14*, the age is categorized in a more fine way into groups of five instead of three. The different age groups were chalked out and the groups are as follows: *a)* 18-24; *b)* 25-34; *c)* 35-49; *d)* 50-64 *e)* 65+. The earlier version sub corpus was distributed in three parts, i.e. training, test and early birds.

2.4.2.1 Social Media

They built the corpus in social media by selecting a part of the *PAN-AP-13* corpus. They chose these authors with an average number of words in their posts above 100. They were also manually reviewed the documents to eliminate those

authors who seem to be fake profiles such as robots, for example, the authors are selling the same product (e.g., laptops, phones) in most of their messages or authors with a large number of reuse text (for example, teenagers sharing poetry or homework). The corpus of social media is balanced by gender, so that the number of sex of the author is half.

2.4.2.2 Blogs

The purpose of the blog is to build a collection gold standard for AP in this specific genre. To achieve this, they selected and manually annotated the documents. First, they sought public LinkedIn profiles that share a personal blog URL. They verified that the blog exists, it is written in one of the languages of Interest (English, Spanish or Dutch) and is updated by one person and that person is easily identifiable. They threw organization blogs when they are not sure that the blog has been updated by the person identified in the LinkedIn profile. Second, they requested information on age. In some cases, the date of birth is published in the user profile. But in most cases it is not that they were seeking the departure date of degree in the education section. They used the information displayed include the age group. They threw Users whose education dates are not clear. Third, if they could include age, they identified gender through photography and name of the user. Again, in cases where information on gender was not clear, they rejected the user. Finally, this process was done by two independent annotators and a third decided in case of disagreement. For each blog, they have provided up to 25 messages. They provided the content obtained from RSS feeds, but they allow users to download the full text of the permalink.

2.4.2.3 Twitter

They have followed the same approach for twitter as earlier done for the blogs. The corpus was developed for the authors in context of their reputation monitoring in twitter. The authors influence among the users was checked in a particular domain. This includes by identifying the category of writers for example journalist, stakeholders, experts and the level of authority of writers on different viewpoints within their fields. It was a tough job to develop the balance list with regard of its age and gender as most of the influential authors were male and were in a narrow range of age from 35-49. The Twitter's corpus is developed and

balanced, half authors are from male category and other half are females.

2.4.2.4 Hotels Reviews

To investigate the applicability of copyright to the nature of the review of profiling methods, they compiled the Webis-TripAd-13 corpus, much of the criticism in *PAN-2014* author profiling evaluation corpus. The corpus was carefully constructed to ensure the quality regarding the cleanliness and accuracy of the annotation text. The Webis-TripAd-13 corpus is derived from another body that was originally used to predict appearance grade level .The original corpus was crawled review TripAdvisor15 website of the hotel in less than a month from mid-February to mid-March 2009 and contains 235,793 comments on 1850 different hotels.

Each review includes the user name of its author, the revised text, and the date the review was written. However, all of the original data is not age and gender annotations. To make this data set applicable to the author of profiling and quality, they applied the following four steps of post-treatment: First, they removed short comments under ten word. Second legal opinion, they removed the text that was not found to be English as a languid historic period detector. Third, since the entire master copy data provides no data on age and gender, they compiled a list of drug user names who submitted comments and crawled the corresponding user profiles from TripAdvisor website. Fourth, since the meta data they rejected all written opinions by writer whose age and gender were not acted on their user profiles, or whose user profile has been inactive.

In addition to ensure data quality, they examined the profiles of users and comments with regard to mental health (i.e., if the selective information given makes sense). The Webis-TripAd-13 final corpus contains 58,101 comments and covers six age classes. To meet the requirements of the authors of *PAN* profiling evaluation corpus, they unify the corpus according to Webis-TripAd-13 for the distribution of age classes of nearly uniform, they sampled 700 authors from all three major classes (25 -34, 35- 49, 50-64). For two secondary classes (18-24, 65+), however, the number of authors available was limited by the size of the smallest age group, so 254 authors (18-24) and 547 authors (65+). 13-17 class was dismissed entirely, as the number of authors available was considered not representative for evaluation [46].

Table 2.2: Statistics of PAN 2014 Corpus

Traits		**Genre**		
		Hotel Reviews	**Blogs**	**Social Media**
Age	18-24	359	6	1550
	25-34	998	60	2098
	35-49	1000	54	2246
	50-64	999	23	1838
	65-xx	799	4	14
Gender	Male	2080	74	3873
	Female	2080	73	3873
$\sum$		4160	147	7746

2.4.3 PAN 2015 Author Profiling Corpus

To investigate how different authors profiling approaches for different languages, they built a corpus of four different languages: English, Dutch, Italian and Spanish. For detection of the age, they followed what has been done before and examined four classes: 18-24, 25-34, 35-49, and 50-xx. In addition to age and gender detection this time as previous 5 most personality traits are considered as outgoing, stable, friendly, conscientious and opened. As regards the character traits for each character is provided scores (between -0.5 and 0.5) [45].

Table 2.3: Statistics of PAN 2015 Corpus

Traits		**No 0f Authors**			
		English	**Dutch**	**Spanish**	**Italian**
Age	18-24	58	-	22	-
	25-34	60	-	46	-
	35-49	22	-	22	-
	50-XX	12	-	10	-
Gender	Male	76	17	50	19
	Female	76	17	50	19
Open	Yes	149	34	83	37
	No	3	0	17	1
Stable	Yes	105	25	53	31
	No	47	9	47	7
Agreeable	Yes	114	26	75	34
	No	38	8	25	4
Extroverted	Yes	120	31	87	30
	No	32	3	13	8
Conscientious	Yes	117	28	77	35
	No	35	6	23	3
$\sum$		152	34	100	38

2.4.4 PAN 2016 Author Profiling Corpus

The main emphasized of 2016 collective task was on cross-genre age and gender detection. The training documents were on one genre (e.g. Twitter) and the evaluation was on another genre (e.g. blogs, social media...). They have worked on only three languages, English, Spanish and Dutch. In *PAN 2016* Corpus they labeled English and Spanish corpora with age and gender but the Dutch corpus was only labeled with gender. They considered the five classes of age: 18-24, 25-34, 35-49,50-64, and 65-xx.

Table 2.4: Statistics of PAN 2016 Corpus

Traits		**No. of Authors**		
		English	**Spanish**	**Dutch**
Age	18-24	26	16	–
	25-34	136	64	–
	35-49	182	126	–
	50-64	78	38	–
	65-xx	6	6	–
Gender	Male	214	125	192
	Female	214	125	192
$\sum$		428	250	384

2.4.5 PAN 2017 Author Profiling Corpus

This year the focus of PAN 2017 concerted task was on cross-genre gender and language variety identification in Twitter. Demographics traits such as gender and language variety have so far investigated separately. In this task they provided a Twitter corpus annotated with authors' gender and their specific variation of their native language:

English (Australia, Canada, Great Britain, Ireland, New Zealand, United States)

Spanish (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela)

Portuguese (Brazil, Portugal) Arabic (Egypt, Gulf, Levantine, Maghrebi)

In our proposed approach we are examining only English language.

Table 2.5: Statistics of PAN 2017 Corpus

Traits		**No. of Authors**
Language (English)	**Australia**	600
	Canada	600
	Great Britain	600
	Ireland	600
	New Zealand	600
	United States	0
Gender	**Male**	1492
	Female	1508
$\sum$		3000

2.4.6 BNC (British National Corpus)

The British National Corpus comprises 920 documents in British English which are marked for finding author gender and for author genre: Many fiction and non-fiction genres were also included. All the experiments shown in the BNC paper were achieved on a genre-controlled subset of the BNC composed. In every subset of genre all the documents selected which were smaller in size, for male and female and the same size of documents were selected from other categories, all the remaining documents were removed. The remaining corpus documents reduced to 566 documents only. Not more than three documents included in the corpus from a single author. The non-fiction and 75% of the fiction documents are from the years 1975-1993; the remaining are from the years 1960-1974. This document includes 61,199 words where the word from female were 34,795 and for male where 33,845 [3].

2.4.7 Formal Text Corpus

The Handbook has a comprehensive analysis of the features for investigating age and gender from the formal text. The handbook also discusses very practical applications of language and gender research in numerous specific localities [36].

2.4.8 Twitter Based Corpus

Sampling of data from Twitter with the use of an API collecting 400,000 tweets per day was initiated in April 2009. Burger *et al.* [9] presented numerous arrangements of a language-independent classifier for forecasting the gender of Twitter users. The bigger dataset made for assessment of these classifiers was drawn from Twitter users who has their blog profile pages.

These were tested on the already gender tagged tweets and the best classifier accuracy was 92% and the other classifier were tested on tweet text only which gave 76% accurate results. The Human performance was only 5% as compared to these classifiers.

2.4.9 Blogs Corpus for Gender Prediction

With the continues growth in the social media, the attention shifted to other kind of writings, more informal, less organized and structured, just like blogs. The dataset on which different experiments was made, contains 566 documents from British National Corpus. The paper reviewed how to address the hitch of automatic detection of an author's gender by recommending simple syntactic and lexical types, and managing to achieve around 80% of accuracy [49].

2.4.10 Weblogs Corpus for Gender and Age prediction

Schler *et al.* [49] reviewed the influence on way of writing used in blogs by age and gender, they collected blogs around 71,000 and introduced a number of stylistic classes e.g. non-dictionary words, parts-of-speech, function words and hyperlinks, merged with information gain words. The results obtained for gender identification were about 80% correct and almost 75% for age identification. They proved the correlations of language writing styles with age.

2.4.11 Segments of Blogs Corpus

Earlier the studies were conducted on the text size of minimum of 250 words and achieved the 80% accuracy for gender predication. However, it was very important

factor for any predictions [38] tested with small segments of blog post, exactly 10,000 segments with 15 tokens per segment, and achieved 72.15 of accuracy.

2.4.12 Blogs and Twitter Corpus

Eemcs *et al.* [34] deemed the utilization of dialect and age among Dutch Twitter clients, in this study the reports were short, with only less than 10 words. They used logistic regression approach and used age as continuous variable. The effect of the gender 22 on different ages were also calculated and the age identified by considering both variables dependent on each other.

2.5 Evaluation Measure

The standard evaluation measure used for AP tasks in recent research is *accuracy* [46, 45] and we also selected it as evaluation measure for our experiments as well. Against each feature the accuracy was calculated by experimenting on corpus. These experiments were applied on supervised data for identification of different demographic traits of the participants/author. For this case both binary and multiple classification was used.

2.5.1 Accuracy

Accuracy is defined as the ratio between total number of correct predictions $n_c$over total number of predictions n_p.

$$Accuracy = \frac{n_c}{n_p}$$

2.6 Chapter Summary

In the current chapter, we explained the overview of the existing methods for AP, existing benchmark corpora. We not only provided background information, the

possible methods to adopt in, the current study also highlights the gap which this study aims to fill. We also discussed Evaluation Measure.

2.7 Summary of Corpora

Here is the table 2.6 for brief summary about other corpora used for AP. Different authors using vast range of features for AP and got some reasonable results on different types of corpora.

Table 2.6: Summary of Existing Corpora

Author	**Data Collection**	**Features**	**Results (accuracy)**
Argamon et al.,	British National Corpus	Part-of-speech	Gender: 0.9
Koppel et al., 2003	Blogs	Simple lexical and syntactic functions	Gender: 0.8
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 0.8, Age: 0.75
Goswami et al., 2009	Blogs	Slang + sentence length	Gender: 0.8918, Age: 0.8032
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/sentence length, POS, word factor analysis	Gender: 0.72
Rangal et al., 2013	Netlog	POS, HTML, n-grams, IR-based, Collection based	Gender: 0.5, Age: 0.33
Rangel et al., 2014	social media, blogs, Twitter, and hotel reviews	Stylistic features, content based features n-grams or bag-of-word	Gender: 0.8846, Age: 0.6923
Rangel et al., 2015	Twitter	stylistic-based and content-based features	Age: 0 .7

Chapter 3

State-of-the-Art for Age and Gender Prediction

State-of-the-art methods in authorship attribution, which aims to regulate an unknown document's author from a set of candidate authors. In this chapter we will describe state-of-the-art approaches of Author Profiling problem for same genre[1] and cross genre[2] respectively. Following is an overview of their feature engineering and summery of their attained results.

3.1 Age and Gender Related Research for Same Genre

Kiprov *et al.* [23] used Lexicon, Twitter-specie, orthographic and Term Level Features and analyzed that most of the orthographic features improving the age and gender accuracy and achieved 84% accuracy for authors gender prediction and more than 70% for age on same genre, by using Support Vector Machine (SVM). *Kiprov et al.* also investigated that sustainable performance among the best feature groups were POS-tag counts, word unigrams and bigrams. Argamon *et al.* [3] tackled the task of gender identification by bringing together function words and parts-of-speech (POS). They noticed properly written texts take out from the British National Corpus and achieved approximately 80% accuracy.

Due to the popularity of electronic media now a days there is a lot of text on social media for that reason social media is the pivot of research, Koppel *et al.* [25] calculated the problem of automatically identify an author's gender by considering the combinations of simple lexical and syntactic features and obtained an accuracy of about 80%. Schler *et al.* [49] examined the effects of age and

[1] Models are trained on one genre, for example trained on Twitter, and evaluated on the same genre but unseen Tweets.

[2] Models are trained on one genre, for example trained on Twitter, and evaluated on another genre different from Twitter.

gender over blog's writing style, the writers collected more than 71,000 blogs and established a set of stylistic features as the words not in the dictionary, parts-of-speech, function words and hyper-links, merged with content based features, such as word unigrams with the highest information gain. They obtained an accuracy of about 80% and 75% to identify gender and age respectively.

Stein *et al.* [54] pointed out that the earlier time studies were handled with at least 250 words of length. The Size of the data set effects the results. Zhang *et al.* [58] carried out trials with short segments of blog post, particularly 10,000 segments with 15 tokens per segment and obtained 72.1% accuracy for gender classification, compared to more than 80% in the previous studies. Stein *et al.* [54] mentioned that most of the participants used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, together with part-of-speech tags and content-based features such as Latent Semantic Analysis, bag of words, TF-IDF, dictionary-based words, topic-based words. Houvardas *et al* and Peersman *et al.* [20, 38] investigated good performance of n-gram features. Holmes [19] worked on the task with 3 million features in a MapReduce configuration and achieved high accuracies with fractions of processing time. Poulston *et al.* [43] have been shown that topic models produced reliable results when used alone and in conjunction with other features.. They have pointed out that n-grams in conjunction with LDA topics is more stable than n-grams on their own, the scores obtained for the English corpora is in the range of 0.7906 to 0.5217.

We can observe that n-grams and topic models are a useful element in developing AP systems across number of languages and providing reasonable results without any additional features. 10-fold cross validation was used in training the data.

3.1.1 Preprocessing Methodology

the main pre-processing step used by Kiprov and Poulston [23, 43] was tokenization. Other authors [17, 21, 33] first place to remove the HTML code from the tweets and handled hashtags, urls and mentions [16, 33, 35, 26].

Gonzalez *et al.* [16] changed mentions, urls, and hashtags for predefined tokens. In a similar way, Maharjan *et al.* [26] substituted the urls with the URL token, or the urls were entirely removed. In spite of the fact that; the dataset was cleaned before publish, Bartoli *et al.* [6] preprocessed tweets to eliminate RTs and shares. Duarteweren *et al.* [12] lowercased the text, removed numbers, stop

words and applied stemming over training datasets. Nowson *et al.* [35] terminated all character sequences depicting emojis in the original tweets, and the authors [41] eliminated tweets with fewer words.

3.1.2 Features

Stein *et al.* [54] reported that many participants approximate the task with different combinations of style-based and content-based feature. Kiprov *et al.* [23, 43] used n-gram models in the composition of style-based and content-based feature. For instance, [26, 16, 55] used character n-grams, [33, 37, 15, 35] worked with word n-grams, [17, 15, 37, 55] used TF-IDF n-grams, whereas [37, 16] took advantage of part of speech n-grams.

With respect to the content-based features [26, 29, 30, 21] used topic modeling with Latent Semantic Analysis (LSA). Maharjan *et al.* [26] played family tokens (my wife/husband, my girlfriend/boyfriend, my hubby, my bf, etc.). The best performing team in *PAN-15* [2], integrated that the Latent Semantic Analysis (LSA) with second order features based on relationships among terms, documents, proles, and sub-proles. Gonzalez *et al.* [16] employed combinations of char and POS n-grams, and [17] combined TF-IDF n-grams with style-based features. Posadas-duran *et al.* [41] presented that syntactic n-grams can be manipulated as features to model author's aspects such as gender and age. They Considered syntactic n-grams as dimensions in a vector space model and used a supervised machine learning approach. They observed that syntactic n-grams of words provided good results when predicting personality traits (RMSE); however, their usage is not that successful when predicting the age and gender.

3.1.3 Results

The following table 3.1 shows the summary of results for English Language AP on same genre.

Table 3.1: Summary of State-of-the-Art Results on Same Genre

Team	**Features**	**Accuracy Achieved**	
		Age	**Gender**
Kiprov et al.[23]	Lexicon, Twitter-Specific, Ortho-graphic, Term Level	0.7	0.8
Argamona et al.[3]	Function words, POS	N/A	0.8
Shler et al.[49]	Stylistic based, Content Based	0.7	0.8
Koppel et al.[25]	Simple Lexicon, Stylistic	N/A	0.8
Zhang et al.[58]	Words, Sentence Length	N/A	0.7
Poulston et al.[43]	N-Grams, LDA Topics	0.5	0.7
Peersman et al.[38]	Token, Character Features	0.7	0.7
Posadas et al. [42]	Syntactic n-grams	0.5	0.5

3.2 Age and Gender Related Research for Cross Genre

In order to study the effect of the cross-genre calculation on the performance of the different author profiling approaches, the researchers used corpora with different genres for training and testing their systems.

3.2.1 Preprocessing Methodology

The pre-processing methodology used in this task is somehow same as mentioned in previous section, although Lemmatization was applied in [8], however the authors expressed no improvement in their results. In [48] the authors tested stem-

ming in pre-processing step. In [8, 14] the writers detached the punctuation signs, stop words were eliminated in [48, 1]. The authors in [8, 1] lowercased the texts and digits were ejected in [8, 27]. Nonetheless, the most familiar pre-processing regarded in Twitter specific components such as hashtags, mentions, RTs or urls [1, 8, 24].

3.2.2 Features

A large number of authors [8, 14, 31, 7, 40] studied different kind of stylistic features. Such as, the frequency of use of function words, words that are not in a predefined dictionary, slang, capital letters, unique words. The adoption of definite sentences per gender (e.g. “my man”, “my wife”, “my girlfriend”...) and age (“I’m” followed by a number) was used in [14] and sentiment words were handled in [14, 40].

The researchers [7, 14, 4, 10] taken into account the parts-of-speech, collocations and LDA [7], different readability indexes [14], [4] used vocabulary richness. [48, 9] modeled the authors with bag-of-words approach. They weighted their n-grams with tf-idf in [1, 11].

3.2.3 Results

The following table 3.2 presents the summary of results for English Language AP on cross genre. They trained their systems on tweets and tested by using Social Media text.

Table 3.2: Summary of State-of-the-Art Results on Cross Genre

Team	**Accuracy Achieved**	
	Age	**Gender**
Busger et al.	0.3046	0.5575
Dichiu & Rancea	0.2989	0.5345
Agrawal & Gonçalves	0.3103	0.5431
Bougiatiotis & Krithara	0.3046	0.5345
Modaresi	0.3218	0.5057
Bilan et al.	0.2902	0.5374
Gencheva et al.	0.2902	0.5287
Kocher & Savoy	0.2816	0.5144
Ashraf et al.	0.2902	0.4971
Bakkar et al.	0.2874	0.5029
Pimas et al.	0.0086	0.0201

3.3 Chapter Summary

In this chapter we explained previous work in the field of Author Profiling, we discussed the corpora used by the researchers for age and gender identification. . This chapter explained state-of-the-art work for cross genre and same genre used by the research community for the task of Author Profiling. We also discussed the existing techniques and measures for the task.

Chapter 4

Proposed Approach

4.1 Introduction

The proposed system has been designed for automatic identification of author's traits particularly age and gender. Our proposed approach is based on Part of Speech tags of syntactic n-grams and traditional n-grams of Part of speech tags, which help us to capture a set of elements of writings. Since male and female are two opposite genders this difference also reflects in their writings, same for different age groups. This natural phenomenon leads us to predict a author's age, gender and other personality traits on the basis of his/her written text. They use different structures, sn-grams of POST and tn-grams of POST help us to capture different individual's structures. The other reason for selecting these features is, in our proposed approach the training data is in one genre and the test data is in another genre. Therefore, these features were expected to accurately identify author traits even if they are trained and tested on different types of data. Figure 4.1 shows the detailed architectural diagram of our work.

In this thesis we propose our approach for author profiling for above mentioned corpora (see section 2). We operated POST based sn-grams and POST based tn-grams analysis for finding the same. The output of each analysis is hand over to the machine learning classifiers which determines the age and gender of the author.

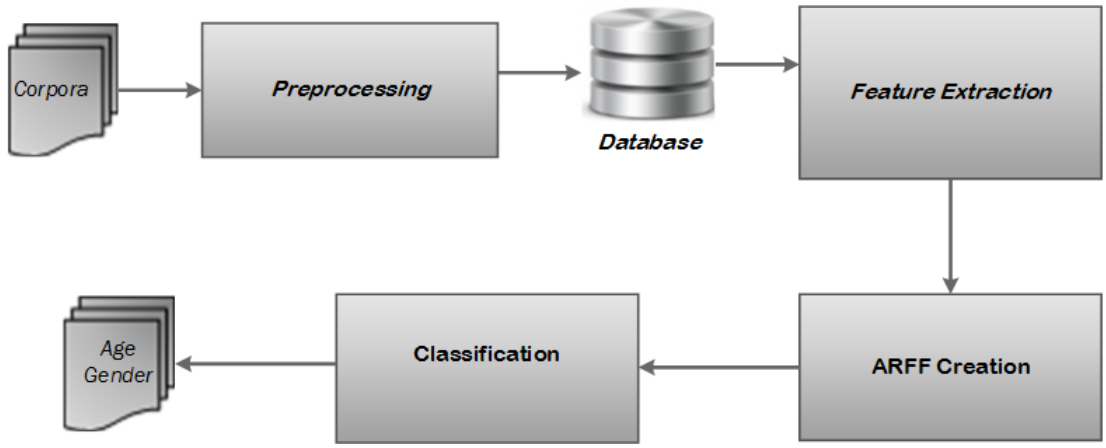

Figure 4.1: System architecture diagram

Figure 4.2 demonstrated main components of the proposed system with their key inputs and outputs.

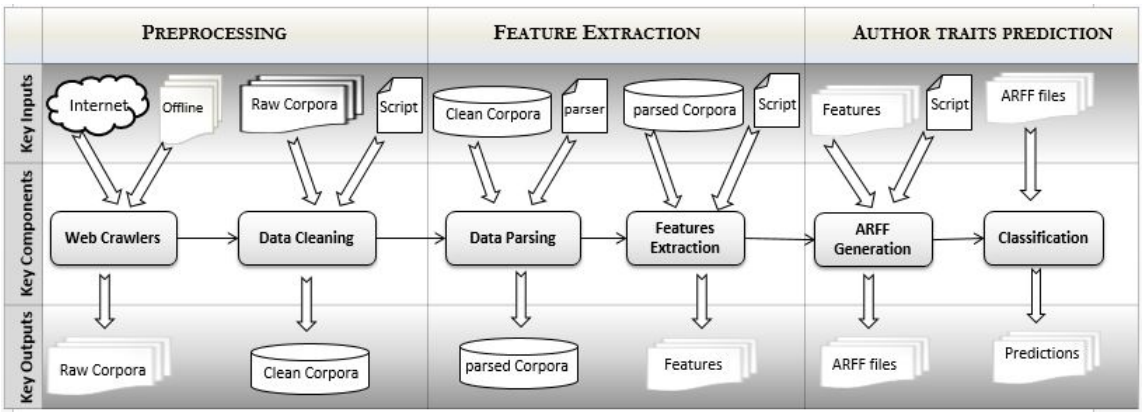

Figure 4.2: Main input/output components of the proposed system

4.2 Proposed Approaches

4.2.1 Preprocessing

For implementation of our approach we used training corpora of *PAN-2014*[1] *and PAN-2016*[2], we considered only English language documents. The data set is labeled with age and gender. The corpora is consisted of different xml documents which had to be handled in an offline (Social media, Blogs and Hotel reviews)

[1]http://pan.webis.de/clef14/pan14-web/author-profiling.html
[2]http://pan.webis.de/clef16/pan16-web/author-profiling.html

and online (tweets) mode. After crawling the data from both modes then cleaned for the removal of xml contents, user mention (twitter), urls, etc and original text written by author is extracted from each xml file and stored in a separate text file for each user. There was no further pre-processing performed on the dataset. The cleaned data is then fed into a database. The detailed procedure of preprocessing is described in 4.3

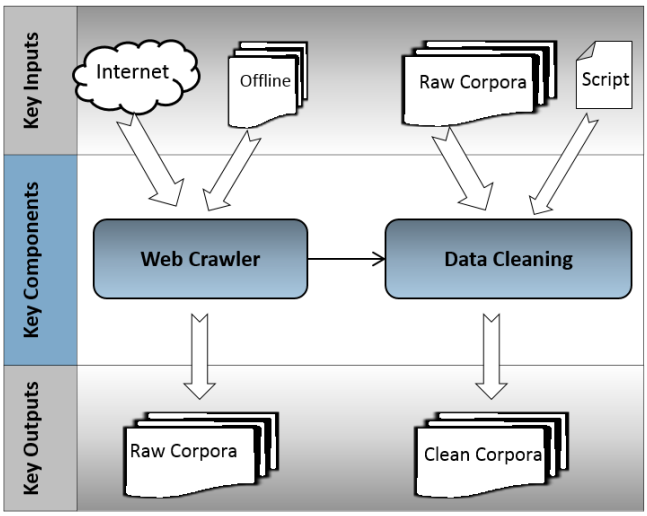

Figure 4.3: Main components of preprocessing stage with their key inputs and outputs

- **Crawlers:** The corpora we are using is provided by *PAN-14 and PAN-16* contains the documents in the XML form. However for some data sets like Twitter and Blogs, data had to be taken from HTML links in the XML file. Hence we had two modes of crawlers; one for offline data sets (Social Media and Hotel Review) and another for online data sets (Twitter and Blogs).

- **Data Cleaning:** The raw text obtained from the crawlers has to be cleaned to remove noisy data like „\ufff, XML tags, urls, twitter user mentions, hashtags etc. The presence of this noisy data could affect and reduce the accuracy of the entire analysis. The cleaned data is then pushed into a database.

After crawling and cleaning step we had following data entries in the database for experiments.

Table 4.1: Gender data set count for PAN-2014 corpora

Corpus Type	**Male/Female**	**Total Count**
Social Media	3873/3873	7746
Blogs	73/74	147
Hotel Review	2080/2080	4160
$\sum$	**6026/6027**	**12053**

Table 4.2: Age group data set count for PAN-2014 corpora

Age Group	**Blogs**	**Hotel Reviews**	**Social Media**	**Total Count**
18-24	06	359	1550	1915
25-34	60	998	2098	3156
35-49	54	1000	2246	3300
50-64	23	999	1838	2860
65 or above	04	799	14	817
$\sum$				**5071**

Table 4.3: Gender data set count for PAN-2016 corpora

Corpus Type	**Male/Female**	**Total Count**
Twitter(tweets)	218/218	436
$\sum$	**218/218**	**436**

Table 4.4: Age group data set count for PAN-2016 corpora

Age Group	**Twitter (tweets)**
18-24	28
25-34	140
35-49	182
50-64	80
65 or above	06
$\sum$	**436**

Table 4.5: Gender group data set count for PAN-2017 corpora

Corpus Type	**Male/Female**	**Total Count**
Twitter(tweets)	1500/1500	3000
$\sum$	**1500/1500**	**3000**

Table 4.6: Language group data set count for PAN-2017 corpora

Traits		**No. of Authors**
Language (English)	**Australia**	600
	Canada	600
	Great Britain	600
	Ireland	600
	New Zealand	600
	United States	0
Gender	**Male**	1492
	Female	1508
$\sum$		3000

We are using the same preprocessing process for all our following proposed approaches.

4.2.2 Traditional N-grams of Part-of-Speech Tags

The term Traditional n-grams of POS tags refers to series of sequential POS tags in sentences, paragraphs and documents. The series can be of length 1 (unigrams), length 2 (bigrams) and length 3 (trigrams) *etc* towards the generalized term n-grams. They represent morphological information and successfully used in various computational linguistic tasks. The sentence in English is following:

Example: ***Iqra reads an interesting book.***

Traditional Bigrams: *Iqra reads, reads an, an interesting, interesting book.*

Traditional Trigrams: *Iqra reads an, reads an interesting, an interesting book.*

Traditional 4-grams: *Iqra reads an interesting, reads an interesting book.*

Traditional 5-grams: *Iqra reads an interesting book.*

Then each word is replaced by its corresponding one of 36 part of speech tags in each sentence by using more correct and robust text analysis tools i.e. Stanford Log-linear Part-Of-Speech Tagger as it gave a 97.2 % accuracy on the Penn Treebank Wall Street Journal corpus [56]. After processing on text by using POS tagger, the following sentence is obtained:

Processed sentence: ***Iqra/NNP reads/VBZ an/DT interesting/JJ book/NN***

Traditional Bigrams of POS Tags: *NNP VBZ*, *VBZ DT*, *DT* JJ, JJ NN.

Traditional Trigrams of POS Tags: NNP VBZ DT, VBZ DT *JJ,* DT *JJ NN.*

Traditional 4-grams of POS Tags: *NNP VBZ DT JJ, VBZ DT JJ NN.*

Traditional 5-grams of POS Tags: *NNP VBZ DT JJ NN.*

The established relation "follow another POS tag" .

4.2.3 Syntactic N-grams of Part-of-Speech Tags

Syntactic n-grams of POS tags is constructed by following path in syntactic trees. Syntactic n-grams represents syntactic information. As though, we still deal with n-grams but avoid the noise by the surface structure of the language due to syntactically unrelated POS tags may appear together. We can gear this anomaly if we follow the actual syntactic relations that link the POS tags even though those tags are not actual neighbors. Let us consider the same example sentence as above:

Example: ***Iqra reads an interesting book.***

In our proposed methodology, we used Stanford Parser[3] for English language example. Figure 4.4 presents the output of the parser as generated by the program.

```
nsubj(reads-2, Iqra-1)
root(ROOT-0, reads-2)
det(book-5, an-3)
amod(book-5, interesting-4)
dobj(reads-2, book-5)
```

Figure 4.4: Stanford parser output

The direct output of Stanford parse is consisted of two parts:

- In first part, the information is presented in terms of constituency grammars.
- In other part, the information is presented in terms of dependency grammars.

The second part of the out is more concerned; in this part, we can see the syntactic dependencies between the pair of words. The parser shows the type of syntactic relation between each pair of the word and their position in the sentence.

Figure 4.5 demonstrating the syntax tree, by using this tree we can directly generate the syntactic n-grams.

[3]Parser is a program that generates syntactic trees. The trees are usually based on formal grammars of various types.

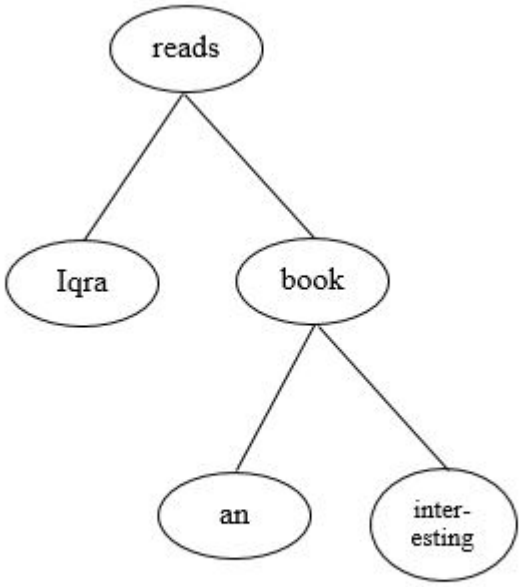

Figure 4.5: Example of syntax tree

Obtained syntactic n-grams are:

Syntactic Bigrams: *reads Iqra, reads book, book an, book interesting .*

Syntactic Trigrams: *reads book an, reads book interesting.*

There are no 4 and 5-grams. Then in next step each word is replaced by its corresponding one of 36 part of speech tags in each sentence by using Stanford Log-linear Part-Of-Speech Tagger and collected Syntactic n-grams of Part-of-Speech Tags are following:

Syntactic Bigrams of POS tags: *VBZ NNP, VBZ NN, NN DT, NN JJ.*

Syntactic Trigrams of POS tags: *VBZ NN DT, VBZ NN JJ.*

There are also other categories of syntactic n-grams depending on the information that adopted for their construction (lemmas, tags of syntactic relations,), all of them are linked through a dependency tree but analyze different linguistic aspects of a sentence. We extracted features from mentioned data set and send it to Naive Bayes, SMO, Logistic, Random Forest and J48 classifiers for training purpose.

4.2.4 Combination of Word and Character Based N-grams

As document is consist of words, and words composed of characters, the form of word/character arrangements can provide valuable information about the content and style of a specific author. So, to achieve better accuracy we used combination of word n-grams (1-3, minimum and maximum grams) and combination of character n-grams (3-5, minimum and maximum grams) as well. (1-3) term means the minimum length is one and maximum is 3 of words, take the first three words of the text and search the unigram, bigram and trigram pairs. (3-5) term refers the minimum length is 3 and maximum is 5 of characters, take the greater than three and less than 5 characters of the text and search the trigram, fourgram and fivegram pair of the characters.

4.3 Chapter Summary

This chapter presenting proposed approaches in detail and work flow of the proposed techniques including traditional n-grams of part-of-speech tags, syntactic n-grams of part-of-speech tags, combination of word and character based n-grams.

Chapter 5

Results and Analysis

5.1 Introduction

In this section, we will report and analyze the results that we have achieved for our experiments using traditional n-grams of part-of-speech tag and and syntactic n-grams of part-of-speech tag based approaches and also the comparison of both approaches.

5.2 Experimental Setup

This section explains the experimental setup used for applying content based methods on above mentioned corpora. (See Section 2.3.3)

5.2.1 Datasets

For the evaluation of the proposed system, three data sets are used. The specifications of these data sets have been shown in chapter 2.

5.2.2 Evaluation Methodology

The problem of identifying an authors' age, gender and language from text is treated as a supervised learning task. Gender identification is a binary classification task because the aim is to distinguish between two classes: (1) male and (2) female. For age identification, we have multi–classification task i.e. goal is

to discriminate between five age groups: *1) 18-24, 2) 25-34, 3) 35-49, 4) 50-64, 5) 65-xx.* For native language, the aim is to distinguish between six classes: *1) Australia, 2) Canada, 3) Great Britain, 4) Ireland, 5) New Zealand, 6) United States.*

We have applied fiive machine learning classifiers, using the WEKA toolkit to find the best classifier for each trait. N-fold cross-validation is used to better estimate the performance of the proposed approach and 10-fold cross-validation is applied on the corpus. The machine learning algorithms we used included J48, Logistic, Random Forest, Support Vector Machines (SMO) and Naive Bayes. we have calculated the percentage of correctly predicted authors' profiles for three traits (PAN-17 corpus for only native language identification task).

5.2.3 Classifiers (Machine Learning Algorithms)

We applied five machine learning classification algorithms as classifiers by using WEKA toolkit to find out the best classifier for each trait. The four classifiers that we used are:

5.2.3.1 Naive Bayes

Naive Bayes [18] is the simplest statistical based classifier. Naive Bayes classifier works on Bayesian rules and takes all the attributes available in data sample and analyzes variables individually independent of each other. The classifier with simple probabilistic properties with no complicated repetitious parameter evaluation mostly performs best as compare to other complicated algorithms.

5.2.3.2 Logistic

Logistic is a statistical method for analyzing a data set in which there are one or more independent variables which determine a result. The result is measured with a dichotomous variable (in which there are only two possible outcomes).

5.2.3.3 SMO

SMO [22] is a supervised machine learning algorithm basically use for classification problems. The need behind the use of this classifier in our experimental setup is to perform classification task for the identification of different demographic traits

5.2.3.4 J48

J48 [44] classifier follows decision tree model for classification of dataset. It develops binary tree for classification process. It decides dependant variable by analyzing all independent variables available in the dataset.

5.2.3.5 Random Forest

Random Forest [44] classifier generates a multiple number of decision trees. Each of the decision tree gives classification for a new object. The random forest then combines the results of all trees for final prediction.

Experiments must carried out using evaluation Measure (see section 2.5).

5.2.4 Features Selection Methods

Raw machine learning data set consists of mix of attributes, some of which are significant to generating predictions. Which features should we use to create a predictive model? We can automatically select those features in our data that are most useful or most suited for the problem we are working on. This procedure called feature selection. Every person has his/her own writing style which can differentiate the person from others. Two things that can vary in a person's written text can be his interests and use of vocabulary in daily routine life as a study say that female witting style is enriched with adverbs and adjectives [53]. As previously discussed that features selection is based on the approach proposed for this work: content model based feature selection.

5.2.4.1 Feature selection algorithm

For our selected content based features, we used only one feature selection algorithm: Info Gain (IG). This feature selection method applies a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset.

5.2.4.2 Information Gain

Information gain (IG) is a quantitative method. It is used for evaluation of effectiveness of features [59]. By using entropy, IG of a feature t can be defined as:

$$IG(t) = H(S) - \sum \frac{|S_v|}{|S|} H(S)$$

IG is informtion gain that we want to calculate for the feature t, H(S) is entropy of dataset S, Sv is subset of dataset S for which feature t hold value v.

5.2.4.3 Benefits of Feature Selection

The feature selection methods are easy to use and give following benefits when used in experiments:

- **Reduce Over fitting:** If the duplication in data or irrelevant data is reduced, we have better chance of making predictions on our data.
- **Improve Accuracy:** On a less ambiguous data we can greatly increase modeling accuracy.
- **Reduce Training Time:** On a minimal size of data, algorithms produce faster results.

5.3 Results and Analysis

5.3.1 Results for Gender Identification

For all tables presented in this section, the following terminologies are used. The "Dataset" is referring to the specific corpus used to evaluate the proposed approaches, corpora including Pan-2014, Pan-2016 and Pan-2017 was manipulated. In Classifier column, we have listed those Machine Learning (ML) algorithms which produced highest accuracy on a given corpora (note that we explored 5 main ML algorithms including Logistic, Random Forest (RF), J48, Naive Bays (NB) and SMO). The "N-gram" describing the number of selected grams chosen for experiments (only mentioned best results). The "Accuracy"defines the accuracy of a classifier for specific features. Table 5.1 shows the best results for gender identification task on same genre using traditional n-grams of part-of-speech tags. Note that we are only reporting the best results on each corpus (for detailed results see Appendix A). Results are produced by using content based method on same genre, same genre means training and testing of the model was on same genre using 10-fold cross validation (CV). Overall, highest accuracy for gender is obtained using content based methods i.e. information gain as feature selection *(accuracy = 0.632653)* with Logistic classifier on Blogs corpus of Pan-2014. For all experiments we are using information gain as feature selection.

Table 5.1: Gender Identification, Best Results using Tn-grams of POST (Same Genre)

Dataset	**Classifier**	**N-gram**	**Accuracy**
Blogs-14	Logistic	3	**0.6326**
HR-14	NB	3	0.6062
SM-14	SMO	4	0.5178
Twitter-16	RF	5	0.5874
Twitter-17	RF	5	0.5083

Table 5.2 presents the best results for gender identification task on cross genre using traditional n-grams of part-of-speech tags. In below table, dataset column refers to the corpora that was used for gender classification experiments. In cross genre model building, training was on one genre and testing of the model was on another genre different to the training corpus. Particularly, in first column of the table "*Blogs to Hotel Reviews*" stated that training of the model was on *Blogs'* data of Pan-2014 and prediction of the trained model was on *Hotel Reviews* of Pan-

2014. We are using Logistic, Random Forest (RF), J48, Naive Bays (NB) and SMO classifiers for content based methods i.e. information gain. The reasonable results *(accuracy = 60.5442)* obtained when training of the system was on *Tweeter Tweets* and evaluation of the model has been done on *Blogs'* data. The accuracy is slightly lower than the same genre's results because most probably features that are used to test the model are different than features that are captured in training the model.

Table 5.2: Gender Identification, Best Results using Tn-grams of POST (Cross Genre)

Training/Testing	**Classifier**	**N-Gram**	**Accuracy**
Blogs to HR	J48	2	0.5233
Blogs to SM	Logistic	1	0.5118
Blogs to Twitter	SMO	1	0.5501
HR to Blogs	Logistic	4	0.5714
HR to SM	SMO	1	0.5086
HR to Twitter	NB	5	0.5338
Twitter to Blogs	NB	4	**0.6054**
Twitter to HR	SMO	1	0.5274
Twitter to SM	Logistic	2	0.5137
SM to Blogs	RF	4	0.5578
SM to HR	RF	3	0.5163
SM to Twitter	Logistic	4	0.5384

Table 5.3 and 5.4 refers to the gender classification results by using syntactic n-grams of part-of-speech tags on same genre and cross genre respectively. The better results are achieved on *Blogs'* data (*accuracy = 0.584615* and *accuracy = 0.561538*) using J48 classifier for same genre and on *Social Media to Blogs* for cross genre experimentation using Logistic classifier respectively. Among the machine learning algorithms, *Logistic* gives highest performance *(accuracy = 0.632653)* using Tn-grams of POST compare to other algorithms for content base approach . After that *Naïve Bayes* gives the best result *(accuracy = 60.5442)* for Tn-grams and *J48* is performing well on Sn-grams for gender identification task.

Table 5.3: Gender Identification, Best Results using Sn-grams of POST (Same Genre)

Dataset	**Classifier**	**N-gram**	**Accuracy**
Blogs-14	J48	4	**0.5846**
HR-14	NB	5	0.5341
SM-14	SMO	5	0.5107
Twitter-16	J48	5	0.5343

Table 5.4: Gender Identification, Best Results using Sn-grams of POST (Cross Genre)

Training/Testing	**Classifier**	**N-gram**	**Accuracy**
Blogs to HR	J48	5	0.5072
Blogs to SM	J48	5	0.5072
Blogs to Twitter	SMO	3	0.493
HR to Blogs	RF	2	0.5538
HR to SM	J48	2	0.5107
HR to Twitter	Logistic	5	0.5121
Twitter to Blogs	Logistic	2	0.5230
Twitter to HR	Logistic	2	0.5233
Twitter to SM	Logistic	5	0.5080
SM to Blogs	Logistic	2	**0.5615**
SM to HR	SMO	3	0.5299
SM to Twitter	NBS	2	0.5062

In the final approach we further tested the combination of two sets of text based features, word n–grams and character n-grams, in the conjunction of SMO classifier to predict age and gender of the producer of the text. We extracted the combination of 1-4 and 3-5 for both word n-grams and character n-grams respectively. In order to estimate the affect of features a 10-fold cross validation was performed on the datasets. Table 5.5 and 5.6 is presenting the results on benchmark corpora of Pan-2014 (Blogs, Hotel Reviews, Social Media) and Pan-2016 (Twitter Tweets) by using Combination of word n-grams and character n-grams respectively. During the experiments it became apparent that the consolidation of word n-grams produced the better results and out perform the start-of-the-art accuracy on same genre of Blogs and Social Media, for Twitter StArt results are NA *(not available)* 5.5 because we are using Tweets corpus of pan-2016 and in that year they were considering cross-genre task of AP. In Table 5.6 best results of gender identification are presented by using combination of character n-grams (3-5) on all four datasets ((Blogs, Hotel Reviews, Social Media and Twitter Tweets) although, the results are not up to the mark by using char n-grams.

Table 5.5: Gender Identification, Best Results using Combination of Word n-grams (Same Genre)

Dataset	**Classifier**	**1-4 w-grams**	**StArt results**
Blogs	SMO	0.7346	0.6795
HR	SMO	0.6659	0.7259
SM[1]	SMO	0.5559	0.5382
Twitter	SMO	0.7062	NA

Table 5.6: Gender Identification, Best Results using Combination of Character n-grams (Same Genre)

Dataset	**Classifier**	**3-5 char-grams**	**StArt results**
Blogs	SMO	0.5918	0.6795
HR	SMO	0.6151	0.7259
SM	SMO	0.5225	0.5382
Twitter	SMO	0.5850	NA

5.3.2 Results for Age Identification

Only the results with higher accuracy were presented for age identification in the table 5.7. After analyzing these results, it has been observed that the outstanding performance of *Naïve Bayes* classifier was achieved by using content based approach on *Blogs-14* corpus *(accuracy = 0.462585)* on same genre. By compare the outcomes on cross genre *(accuracy = 40.8163)* the results are negligibly less than the results on same genre that is reasonable due to the different nature of the corpora in training and testing the model.

Table 5.7: Age Identification, Best Results using Tn-grams of POST (Same Genre)

Dataset	**Classifier**	**N-gram**	**Accuracy**
Blogs-14	NB	1	**0.4625**
HR-14	RF	2	0.2712
SM-14	RF	5	0.3329
Twitter-16	NB	3	0.4209

Table 5.8: Age Identification, Best Results using Tn-grams of POST (Cross Genre)

Training/Testing	**Classifier**	**N-gram**	**Accuracy**
Blogs to HR	Logistic	2	0.2495
Blogs to SM	J48	3	0.3050
Blogs to Twitter	SMO	3	0.3201
HR to Blogs	NB	3	0.3537
HR to SM	NB	2	0.2939
HR to Twitter	NB	5	0.3338
Twitter to Blogs	NB	2	**0.4081**
Twitter to HR	SMO	1	0.2486
Twitter to SM	RF	1	0.2993
SM to Blogs	NB	2	0.3673
SM to HR	RF	3	0.3333
SM to Twitter	Logistic	3	0.3584

Table 5.9 and 5.10 refers to the results for age classification task by using sn-grams of POST on both same genre and cross genre respectively. If we compare the results of age attribute for both tn-grams and sn-grams of POST the accuracy is same by using sn-grams on same genre as we achieved through tn-grams of POST *(accuracy = 0.462585)*. On cross genre traditional n-grams of part-of-speech tags are performing slightly better than the syntactic n-grams of part-of-speech tags with minor difference.

Table 5.9: Age Identification, Best Results using Sn-grams of POST (Same Genre)

Dataset	**Classifier**	**N-gram**	**Accuracy**
Blogs-14	NB	1	**0.4625**
HR-14	RF	2	0.2712
SM-14	SMO	5	0.3329
Twitter-16	NB	3	0.4209

Table 5.10: Age Identification, Best Results using Sn-grams of POST (Cross Genre)

Training/Testing	**Classifier**	**N-gram**	**Accuracy**
Blogs to HR	Logistic	2	0.2495
Blogs to SM	J48	3	0.3050
Blogs to Twitter	Logistic	2	**0.3968**
HR to Blogs	NB	4	0.3537
HR to SM	NB	3	0.2911
HR to Twitter	Logistic	3	0.3968
Twitter to Blogs	SMO	4	0.3615
Twitter to HR	SMO	2	0.2486
Twitter to SM	SMO	3	0.2937
SM to Blogs	NB	2	0.3673
SM to HR	RF	3	0.3333
SM to Twitter	J48	4	0.3968

Table 5.11 and 5.12 are displaying the best results for age attribute by using combination of word based n-grams (1-4) and character based n-grams (3-5). It can be seen that wn-gram approach out performed the StArt results on Blogs and Social Media corpora, char n-grams also performed well just on blogs corpus.

Table 5.11: Age Identification, Best Results using Combination of Word n-grams (Same Genre)

Dataset	**Classifier**	**1-4 w-grams**	**StArt results**
Blogs	SMO	0.4965	0.3974
HR	SMO	0.3217	0.3502
SM	SMO	0.3892	0.3652
Twitter	SMO	0.4731	NA

Table 5.12: Age Identification, Best Results using Combination of Character n-grams (Same Genre)

Dataset	**Classifier**	**3-5 char-grams**	**StArt results**
Blogs	SMO	0.4013	0.3974
HR	SMO	0.2618	0.3502
SM	SMO	0.2843	0.3652
Twitter	SMO	0.4289	NA

We used two types of features as baseline for comparison purposes: word based features and character based features. For baseline features, we used combinations of words (1-4) and character (3-5) n-gram technique. In Tables 5.13 to 5.16 the results of baseline methods are presented for both attributes gender and age. For better appreciation of the comparison of the results, we present Tables 5.17 and 5.18 for gender and age respectively. SMO was giving better results out of all other tried classifiers, so for experimentation of word based and character based combinations we only mentioned accuracy of SMO classifier and also just SMO's results are presented for the baseline.

Table 5.13: Word Based Uni-grams (Gender Baseline).

Classifiers	**corpora results**			
	Blogs	**HR**	**SM**	**Twitter**
NB	0.6258	0.6737	0.6667	0.6083
Logistic	0.6802	0.6737	0.6597	0.6223
SMO	0.6734	0.6612	0.6743	0.6932
J48	0.6462	0.6106	0.6186	0.6270
RM	0.6906	0.6673	0.6692	0.7042

Table 5.14: Word Based Uni-grams (Age Baseline).

Classifiers	**corpora results**			
	Blogs	**HR**	**SM**	**Twitter**
NB	0.4184	0.2849	0.3343	0.3496
Logistic	0.4297	0.2777	0.3482	0.3573
SMO	0.4064	0.2994	0.3675	0.3354
J48	0.4326	0.2811	0.3578	0.3449
RM	0.4113	0.2900	0.3836	0.4648

Table 5.15: Character Based 3-grams (Gender Baseline).

Classifiers	**corpora results**			
	Blogs	**HR**	**SM**	**Twitter**
NB	0.6394	0.6343	0.6203	0.5897
Logistic	0.6258	0.6206	0.6367	0.5664
SMO	0.6190	0.5656	0.5982	0.6363
J48	0.4625	0.5778	0.5578	0.6293
RM	0.6462	0.5496	0.5396	0.6806

Table 5.16: Character Based 3-grams (Age Baseline).

Classifiers	**corpora results**			
	Blogs	**HR**	**SM**	**Twitter**
NB	0.3741	0.2656	0.2786	0.3566
Logistic	0.4184	0.2864	0.2923	0.3664
SMO	0.4539	0.2956	0.2999	0.3363
J48	0.3900	0.2317	0.2422	0.3100
RM	0.4042	0.2827	0.2950	0.4638

In the following Tables we presented the results obtained by using combination of word and character based n-grams along with baseline results in order to compare with the selected methods of baseline.

Table 5.17: Comparison of Word and Character Based n-grams with Baseline Results (gender)

Dataset	**Features**			
	combination of wn-grams (1-4)	**combination of char n-grams (3-5)**	**word uni-gram** (baseline)	**char 3-gram** (baseline)
Blogs	0.7346	0.5918	0.6734	0.6612
HR	0.6659	0.6151	0.6612	0.5656
SM	0.5559	0.5225	0.6743	0.6932
Twitter	0.7062	0.5850	0.6932	0.6363

Table 5.18: Comparison of Word and Character Based n-grams with Baseline Results (age)

Dataset	**Features**			
	combination of wn-grams (1-4)	**combination of char n-grams (3-5)**	**word uni-gram** (baseline)	**char 3-gram** (baseline)
Blogs	0.4965	0.4013	0.4064	0.4539
HR	0.3217	0.2618	0.2994	0.2956
SM	0.3892	0.2843	0.3675	0.2999
Twitter	0.4731	0.4289	0.3635	0.3363

Results are only as good as someone has data. It is critical that we feed machine learning algorithms the right data for the problem we want to solve.

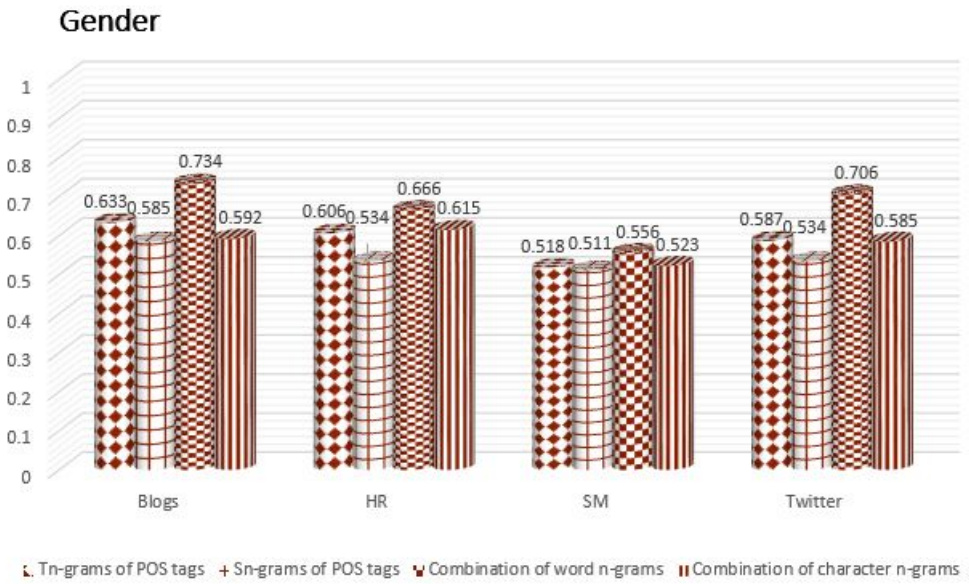

Figure 5.1: Comparison of best results for gender (same genre)

To summarize, Fig. 5.1 presents the utmost results obtained for gender, among all the approaches applied on mentioned corpora. The highest accuracy 0.734 is achieved on Blogs-14 corpus with combination of word-ngrams (1, 2, 3, 4) and it is comparable to the best result obtained on blogs corpus (English) in PAN-2014 Author Profiling Competitions.

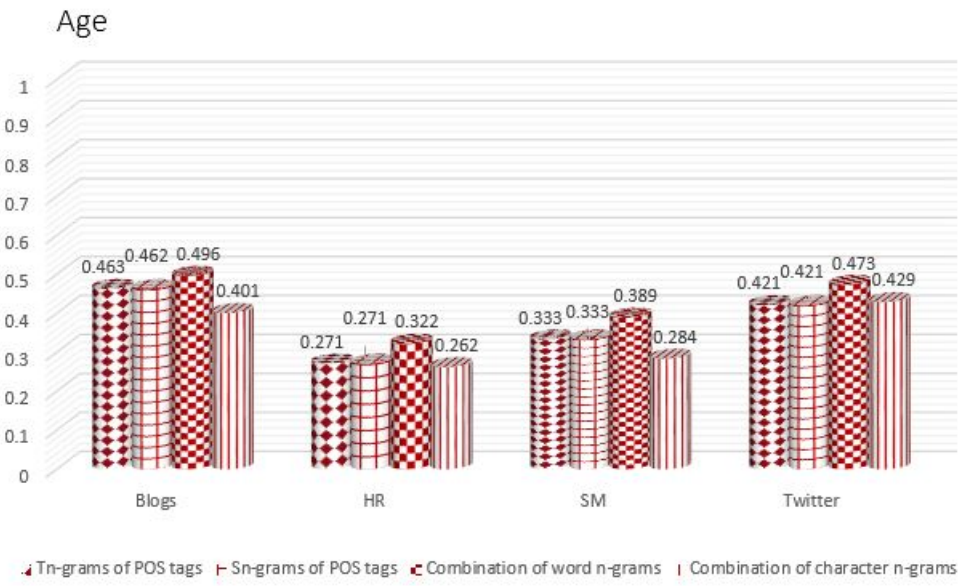

Figure 5.2: Comparison of best results for Age (same genre)

Fig. 5.2 shows the best results obtained for age, among all the approaches applied on mentioned corpora. The highest accuracy 0.493 is achieved on Twitter-16 corpus with combination of word-ngrams (1, 2, 3, 4). It can be appreciated that for gender, combination of word based n-gram technique surpass the baseline techniques for all the datasets except Social Media corpus, for age it outperformed all the cases. Combination of character n-grams are not performing well for gender however on age, just for twitter corpus.

To conclude, inclusively the word based features outperform other features for both age and gender identification tasks. The conceivable reasoning is, communication on blogs and twitter is usually formal and word based features are likely to capture more discriminative information from the text.

5.3.3 Result's summary for Gender Identification (Same Genre)

Table 5.19: Result's summary for Gender identification (Same Genre)

Dataset	Features					StArt
	Tn-gram (POS)	**Sn-gram (POS)**	**COW (1-4)**	**COC (3-5)**	**word uni-gram** (baseline)	**char 3-gram** (baseline)	
Blogs	63	58	**73**	59	67	66	67
HR	60	53	66	61	66	56	72
SM	51	51	55	52	67	69	53
Twitter	58	53	70	58	69	63	NA

5.3.4 Result's summary for Age Identification (Same Genre)

Table 5.20: Result's summary for Age identification (Same Genre)

Dataset	Features						StArt
	Tn-gram (POS)	**Sn-gram (POS)**	**COW (1-4)**	**COC (3-5)**	**word uni-gram** (baseline)	**char 3-gram** (baseline)	
Blogs	46	46	**49**	40	40	45	39
HR	27	27	32	26	29	29	35
SM	33	33	38	28	36	29	36
Twitter	42	42	47	42	36	33	NA

It is appreciable that our system surpassed the state-of-the-art and baseline methods for same genre as we compared and showed in table 5.19 and 5.20.

5.4 Chapter Summary

This chapter discussed about experimental setup including data set used for experiments, evaluation methodology, feature selection methods. Also described in detail about machine learning tools and classification approaches used in this research. It is also covering an overview of evaluation measures used. This chapter also presenting the results and detailed analysis of the study.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The specific goals are achieved:

- Collected a benchmark corpora from different genres (Social Media, Hotel Reviews, Twitter and Blogs) for cross genre prediction of authors's age and gender.

- Preprocessed the collected corpora and cleaned it after removing tags e.g. HTML tags, URLs.

- Extracted the features for preprocessed corpora. We extracted Syntactic and Traditional n-grams of Part-of-Speech tags, Word and character n-grams.

- Designed the experiments for preprocessed corpora.

- Performed the experiments on benchmark corpora in order to predict author's age and gender.

- Evaluated the system on four benchmark corpora from different genres e.g. Social Media, Hotel Reviews, Twitter and Blogs.

In this research work we have discussed the Author Profiling task. For this purpose, we collected the benchmark corpora of PAN-2014, 2106 and used in our research (see section 2.4.2 and 2.4.4). We preprocessed the collected corpora and cleaned it after removing tags e.g. HTML tags, URLs to get meaningful infor-

mation out of the text.we mainly used syntactic n-grams of part-of-speech tags, traditional n-grams of part-of-speech tags, combination of word based n-grams and combination of character based n-grams. The main contribution of the approach is that syntactic and traditional n-grams of part-of-speech tags can be used as features to make prediction about author's attributes such as gender and age. It is possible to tackle the Author Profiling problem after considering syntactic n-grams of part-of-speech tags and traditional n-grams of part-of-speech tags as dimensions in a vector space model and using a supervised machine learning approach. We conducted experiments for author profiling task using various Machine Learning algorithms including Naive Bayes, Logistic, SMO, J48 and Random Forest but best results are achieved using combination of word based n-grams with SMO on blogs and twitter datasets. We used word uni-grams and character 3-grams as baseline. The results showed that combination of word-grams (1-4) technique outperforms the baseline technique and some of state-of-the-art results on same corpora, combinations of character based n-grams (3-5) are not performing up to the mark .

6.2 Final Contribution

Following are the final technical contributions of the study:

6.2.1 Final Technical Contributions

Following are our final technical contributions:

- Creation of dependency trees for preprocessed corpora for preprocessed corpora.
- Extraction of Traditional N-grams of Part of Speech tags for preprocessed corpora.
- Extraction of Syntactic N-grams of Part of Speech tags for preprocessed corpora.

6.2.2 Final Scientific Contributions

We achieved following scientific contributions:

- Comparison of different machine learning algorithms for author's age and gender for various corpora and for cross genre conditions .

- Comparison of various feature sets on a range of benchmark author profiling corpora on different genres including Social Media, Hotel Reviews, Twitter and Blogs.

- Comparison of results for writer's age and gender with baseline and state-of-the-art results on same corpora.

As our proposed approach (Tn-grams of POST and Sn-grams of POST) attaining the information contained in the dependency trees and by using Stanford part-of-speech tagger, the performance is influenced by the use of external resources i.e syntactic parsers and Stanford part-of-speech tagger. Despite the fact, most of the parsers have recently encountered important enhancements, they still facing the certain problems concerning the noise data analysis. The addition of the noise occurred by the usage of external tools, and this is one of the reasons why the approach Tn-grams of POST and Sn-grams of POST did not show very good results on benchmark corpora.

6.3 Future Work

The promising avenues of future research work are: investigate more attributes like native language, profession, level of education, living city etc. Integration with Neural Network Deep Learning can be an interesting direction. Increase the size of the corpora. Exploring other techniques for author profiling can be another avenue.

Bibliography

[1] Madhulika Agrawal and Teresa Gonçalves. Age and Gender Identification using Stacking for Classification—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[2] Miguel A. Álvarez-Carmona, A. Pastor López-Monroy, Manuel Montes y Gómez, Luis Villaseñor-Pineda, and Hugo Jair Escalante. INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[3] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-*, 23(3):321–346, 2003.

[4] Shaina Ashraf, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. Cross-Genre Author Profile Prediction Using Stylometry-Based Approach—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[5] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[6] Alberto Bartoli, Andrea De Lorenzo, Alessandra Laderchi, Eric Medvet, and Fabiano Tarlao. An author profiling approach based on language-dependent content and stylometric features. In *Proceedings of CLEF*, 2015.

[7] Ivan Bilan and Desislava Zhekova. CAPS: A Cross-genre Author Profiling System—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[8] Konstantinos Bougiatiotis and Anastasia Krithara. Author Profiling using Complementary Second Order Attributes and Stylometric Features—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[9] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[10] Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[11] Daniel Dichiu and Irina Rancea. Using machine learning algorithms for author profiling in social media. *Balog et al.[5]*.

[12] Edson Roberto Duarte Weren. Information Retrieval Features for Personality Traits—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[13] Michael Fire, Dima Kagan, Aviad Elyashar, and Yuval Elovici. Friend or foe? fake profile identification in online social networks. *Social Network Analysis and Mining*, 4(1):1–23, 2014.

[14] Pepa Gencheva, Martin Boyanov, Elena Deneva, Preslav Nakov, Yasen

Kiprov, Ivan Koychev, , and Georgi Georgiev. PANcakes Team: A Composite System of Genre-Agnostic Features For Author Profiling—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[15] Maite Giménez, Delia Irazú Hernández, and Ferran Pla. Segmenting Target Audiences: Automatic Author Profiling Using Tweets—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[16] Carlos E González-Gallardo, Azucena Montes, Gerardo Sierra, J Antonio Núñez-Juárez, Adolfo Jonathan Salinas-López, and Juan Ek. Tweets classification using corpus dependent tags, character and pos n-grams. In *Proceedings of CLEF*, 2015.

[17] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. Author profiling using stylometric and structural feature groupings. In *Proceedings of CLEF*, 2015.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[19] Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*, volume 25. John Wiley & Sons, 2008.

[20] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer, 2006.

[21] Hafiz Rizwan Iqbal, Muhammad Adnan Ashraf, and Rao Muhammad Adeel Nawab. Predicting an author's demographics from text using topic modeling approach. In *Proceedings of CLEF*, 2015.

[22] S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and

Karuturi Radha Krishna Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural computation*, 13(3):637–649, 2001.

[23] Yasen Kiprov, Momchil Hardalov, Preslav Nakov, and Ivan Koychev. SU-PAN'2015: Experiments in Author Profiling—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[24] Mirco Kocher and Jacques Savoy. UniNE at CLEF 2016: Author Profiling—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[25] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[26] Suraj Maharjan and Thamar Solorio. Using wide range of features for author profiling. In *Proceedings of CLEF*, 2015.

[27] Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. Adapting Cross-Genre Author Profiling to Language and Corpus—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[28] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. In *Proceedings of CLEF 2014 Evaluation Labs*, pages 1129–1136, 2014.

[29] Caitlin McCollister, Shu Huang, and Bo Luo. Building Topic Models to Predict Author Attributes from Twitter Messages—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[30] Lesly Miculicich Werlen. Statistical Learning Methods for Profiling Analysis—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[31] Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad. Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[32] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[33] Fahad Najib, Waqas Arshad Cheema, and Rao Muhammad Adeel Nawab. Author's traits prediction on twitter data using content based approach. In *Proceedings of CLEF*, 2015.

[34] D-P. Nguyen and A. S. Dogruoz. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA*, pages 857–862, Michigan, USA, October 2013. Association for Computational Linguistics.

[35] Scott Nowson, Julien Perez, Caroline Brun, Shachar Mirkin, and Claude Roux. Xrce personal language analytics engine for multilingual author profiling. *Working Notes Papers of the CLEF*, 2015.

[36] Anne O'Keeffe and Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010.

[37] Alonso Palomino-Garibay, Adolfo T. Camacho-González, Ricardo A. Fierro-Villaneda, Irazú Hernández-Farias, Davide Buscaldi, and Ivan V. Meza-Ruiz. A Random Forest Approach for Authorship Profiling—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San

Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[38] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[39] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[40] Oliver Pimas, Andi Rexha, Mark Kröll, and Roman Kern. Profiling Microblog Authors Using Concreteness and Sentiment—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[41] Juan-Pablo Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas. Syntactic N-grams as Features for the Author Profiling Task—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[42] Juan-Pablo Posadas-Durán, Grigori Sidorov, Ildar Batyrshin, and Elibeth Mirasol-Meléndez. Author Verification Using Syntactic N-grams—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[43] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Topic Models and n-gram Language Models for Author Profiling—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[44] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[45] Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[46] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd Author Profiling Task at PAN 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org, September 2014.

[47] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the Author Profiling Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, September 2013.

[48] José Duarte Rodwan Bakkar Deyab and Teresa Gonçalves. Author Profiling Using Support Vector Machines—Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org, September 2016.

[49] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

[50] Grigori Sidorov. Non-linear construction of n-grams in computational linguistics. *México: Sociedad Mexicana de Inteligencia Artificial*, 2013.

[51] Grigori Sidorov. Syntactic dependency based n-grams in rule based automatic english as second language grammar correction. *International Journal of Computational Linguistics and Applications*, 4(2):169–188, 2013.

[52] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860, 2014.

[53] Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[54] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. Overview of the PAN/CLEF 2015 Evaluation Lab. In Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J.F. Jones, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Initiative (CLEF 15)*, pages 518–538, Berlin Heidelberg New York, September 2015. Springer.

[55] Octavia-Maria Șulea and Daniel Dichiu. Automatic Profiling of Twitter Users Based on Their Tweets—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September 2015.

[56] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[57] G Udny Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390, 1939.

[58] Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010.

[59] Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting bleu/nist scores:

How much improvement do we need to have a better system? In *LREC*, 2004.

Appendix A

Appendix

A.1 Results for Gender Identification

A.2 Gender Traditional n-grams of Part-of-Speech Tags (Same Genre)

Table A.1: Blogs-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	52.381%	57.8231 %	58.5034%	57.8231%	48.9796%
Logistic	50.3401	59.8639%	**63.2653%**	**59.1837%**	51.7007%
SMO	54.4218%	59.1837 %	58.5034%	55.7823%	51.7007%
J48	**55.7823%**	56.4626 %	56.4626%	53.7415%	**61.2245%**
RF	53.7415%	**61.9048 %**	59.8639%	58.5034%	59.8639%

Table A.2: Hotel Reviews-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	1	2	3	4	5
NB	53.6779%	**58.3654%**	**60.6250%**	**59.7115%**	**59.3269%**
Logistic	**54.5673%**	55.1923%	56.1298%	55.240%	53.9904%
SMO	53.5337%	56.7548%	55.9619%	55.8413%	53.6058%
J48	50.2404%	51.5865%	52.8125%	50.8413%	52.5721%
RF	49.6875%	57.6442%	52.6442%	56.7788%	52.5721%

Table A.3: Social Media-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	51.3814%	50.9295%	49.8322%	49.8322%	50.3873%
Logistic	51.5492%	51.4717%	**51.5492%**	50.3098%	50.9424%
SMO	**51.7041%**	51.5879%	50.1010%	**51.7823%**	51.0007%
J48	51.2264%	50.5035%	50.6971%	50.2324%	50.1420%
RF	50.2453%	**51.6266%**	50.5293%	50.3098%	**51.3426%**

Table A.4: Twitter-16 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	48.7179%	**52.9138%**	49.1841%	49.4172%	49.8834%
Logistic	46.8531%	53.3800%	48.0186%	**51.049%**	54.7786%
SMO	48.9510%	52.4476%	51.0490%	50.5828%	55.0117%
J48	47.3193%	48.7179%	**51.0490%**	50.1166%	48.9510%
RF	**53.3800%**	51.7483%	50.8159%	50.5828%	**58.7413%**

A.3 Gender Traditional n-grams of Part-of-Speech Tags (Cross Genre)

Training on Blogs and Testing on Hotel reviews.

Table A.5: Blogs To Hotel Reviews-14 Gender Identification, Results using Tn-grams of POST on (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	50.6731%	50.3365%	50.3365%	**50.5288%**	**50.6250%**
Logistic	**51.5625%**	50.1683 %	50.1683 %	48.9183%	49.9279%
SMO	51.3702%	50.1202 %	50.1202 %	49.8317%	49.7115%
J48	50.8413%	**52.3317%**	**52.3317%**	49.6875%	49.5192%
RF	51.0817%	51.0817%	51.7788%	48.6779%	49.5673%

Table A.6: Blogs To Social Media-14 Gender Identification, Results using Tn-grams of POST on (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	50.0645%	49.3932%	49.3416%	50.1166%	49.8838%
Logistic	**51.1877%**	50.3873%	48.9801%	50.284%	50.2066%
SMO	50.2195%	**50.6971%**	50.284%	**50.4777%**	50.142%
J48	50.5551%	50.4648%	**50.5939%**	50.1807%	48.9543%
RF	49.9225%	50.4389%	49.7418%	50.0516%	**50.2969%**

Table A.7: Blogs To Twitter-16 Gender Identification, Results using Tn-grams of POST on (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	50.1166%	50.1166%	49.6503%	48.9510%	50.1166%
Logistic	52.2145%	49.1841%	50.3497%	48.0186%	50.3497%
SMO	**55.0117**	**52.4476%**	48.2517%	48.2517%	50.5828%
J48	52.2145%	48.4848%	**50.3497%**	**50.1166%**	48.0186%
RF	51.9814%	49.4172%	48.7179%	45.6876%	**51.7483%**

Table A.8: Hotel Reviews To Blogs-14 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	51.7007%	49.6599%	51.7007%	48.2993%	47.6190%
Logistic	51.7007%	**56.4626 %**	**55.1020%**	57.1429 %	51.0204
SMO	**53.7415%**	53.7415 %	53.0612%	**57.1429%**	**53.0612%**
J48	49.6599%	54.4218%	48.2993%	53.7415%	44.898%
RF	51.7007%	53.0612%	51.7007%	47.6190%	50.3401%

Table A.9: Hotel Reviews To Social Media-14 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	50.1678%	49.8709%	49.8193%	49.8580%	**50.1678%**
Logistic	50.3486%	49.5740%	49.8580%	50.1936%	49.7418%
SMO	**50.865%**	**50.5293%**	**50.3615%**	50.1420%	49.6127%
J48	49.4965%	48.9027%	49.7160%	**50.1936%**	49.3416%
RF	49.8838%	49.8967%	49.9355%	50.0258%	49.8967%

Table A.10: Hotel Reviews To Twitter-16 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	**52.2145%**	51.0490%	**51.2821%**	51.7483%	**53.3800%**
Logistic	51.0490%	**51.2821%**	51.2821%	52.2145%	52.2145%
SMO	50.4087%	51.0490%	51.0490%	**53.1469%**	49.8834%
J48	47.7855%	49.6503%	50.8159%	52.2145%	49.4172%
RF	49.1841%	51.0490%	48.4848%	51.7483%	49.4172%

Table A.11: Twitter To Blogs-14 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	51.0204%	**58.5034%**	**57.8231%**	**60.5442%**	**57.8231%**
Logistic	48.2993%	56.4626%	52.3810%	57.8231%	51.7007%
SMO	48.9796%	53.0612%	53.7415%	56.4626%	53.7415%
J48	**57.8231%**	53.7415%	47.6190%	50.3401%	55.7823%
RF	53.7415%	57.8231%	56.4626%	52.3810%	48.9796%

Table A.12: Social Media To Blogs-14 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	42.1769%	49.6599%	48.2993%	**52.3810%**	52.3810%
Logistic	55.1020%	**52.3810%**	47.6190%	51.7007%	47.6190%
SMO	**55.1020%**	49.2125%	48.3863%	49.6256%	49.7676%
J48	48.9796%	49.6599%	46.9388%	48.2993%	**53.0612%**
RF	48.9796%	49.6599%	**55.7823%**	52.3810%	51.0204%

Table A.13: Social Media To Hotel Reviews-14 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	49.7356%	49.9038%	49.8798%	49.8798%	50.2885%
Logistic	50.0721%	50.9615%	51.2500%	49.9279%	**50.3846%**
SMO	50.7933%	49.2125%	48.3863%	49.6256%	49.7676%
J48	50.3365%	**51.5865%**	48.2212%	50.9615%	50.2885%
RF	**50.8654%**	50.5288%	**51.5385%**	**51.6346%**	49.3510%

Table A.14: Social Media To Twitter-16 Gender Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	**51.7483%**	48.7179%	48.7179%	48.2517%	48.0186%
Logistic	49.6503%	45.4545%	46.8531%	48.951%	48.9510%
SMO	46.8531%	46.8531%	46.8531%	49.6256%	49.7676%
J48	48.0186%	**50.8159%**	**49.4172%**	47.7855%	**53.8462%**
RF	49.8834%	46.3869%	46.3869%	**49.6503%**	47.0862%

A.4 Gender Syntactic n-grams of Part-of-Speech Tags (Same Genre)

Table A.15: Blogs-14 Gender Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**53.8462 %**	**53.8462%**	52.3077%	46.1538%
Logistic	51.5385%	43.0769%	46.9231%	46.1963%
SMO	50 %	52.3077%	50.7692%	46.1538%
J48	45.3846%	46.1538%	**58.4615%**	**47.6923%**
RF	47.6923 %	45.3846%	45.3846%	46.9231%

Table A.16: Hotel Reviews-14 Gender Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**52.7097 %**	**53.4132%**	**52.1105%**	**53.0769%**
Logistic	52.0323%	52.0769%	50.4615%	51.3288%
SMO	51.8499%	52.1626%	51.3288%	51.0769%
J48	51.4851%	49.6613%	51.8462%	51.9231%
RF	51.3288%	51.3288%	50.0000%	50.9231%

Table A.17: Social Media-14 Gender Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	49.4417%	48.9483%	49.5456%	49.0769%
Logistic	49.6494%	49.0769%	49.0769%	50.3288%
SMO	48.3251%	49.0782%	49.1041%	**51.0769%**
J48	49.8312%	**50.1169%**	**50.8462%**	46.9231%
RF	**50.818%**	50.1088%	50.0000%	50.9231%

Table A.18: Twitter-16 Gender Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	51.5625%	48.4375%	44.6875%	46.5625%
Logistic	**51.5625%**	51.1625%	**51.2625%**	51.4625%
SMO	51.5625%	**53.125%**	49.0625%	50.0000%
J48	45.3125%	51.2500%	49.0625%	**53.4375%**
RF	47.8125%	50.6250%	50.625%	50.6250%

A.5 Gender Syntactic n-grams of Part-of-Speech Tags (Cross Genre)

Table A.19: Blogs To Hotel Reviews-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	49.5310%	49.2705%	**49.2183%**	48.9057%
Logistic	**49.7655%**	**50.4429%**	48.3846%	48.9279%
SMO	48.5409%	50.3387%	48.9578%	49.1402%
J48	49.7655%	48.2022%	49.0620%	**50.7295%**
RF	49.6092%	49.2965%	49.0620%	50.3908%

Table A.20: Blogs To Social Media-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.0909%	49.8052%	50.1688%	49.8312%
Logistic	49.6235%	48.2991%	**48.8185%**	48.2066%
SMO	49.9610%	**49.2340%**	49.5456%	49.9351%
J48	49.4677%	49.3378%	49.4677%	**50.7295%**
RF	51.3373%	50.1948%	50.0909%	50.6102%

Table A.21: Blogs To Twitter-16 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	49.3750%	49.3750%	49.3750%	49.3750%
Logistic	**49.3750%**	49.375%	**49.3750%**	49.3750%
SMO	49.3750%	**49.375%**	49.3750%	49.3750%
J48	49.3750%	49.375%	49.3750%	**49.3750%**
RF	49.3750%	49.375%	49.3750%	49.3750%

Table A.22: Hotel Reviews To Blogs-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.7692%	**53.8462%**	46.9231%	47.6190%
Logistic	49.2308%	49.4626 %	51.1429%	51.0204%
SMO	48.4615%	50.7692%	**53.0769%**	**53.0612%**
J48	46.9231%	50.4218%	50.0000%	44.8980%
RF	**55.3846%**	52.0612%	50.0000%	50.3401%

Table A.23: Hotel Reviews To Social Media-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	49.1820%	**50.0909%**	**50.6881%**	47.6190%
Logistic	**50.5583%**	49.8580%	50.1936%	51.0204%
SMO	49.0782%	49.7273%	49.8052%	**53.0612%**
J48	51.0776%	49.7160%	49.4157%	44.8980%
RF	50.5323%	49.9355%	49.3119%	50.3401%

Table A.24: Hotel Reviews To Blogs-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.6250%	**50.6250%**	**50.6250%**	50.3800%
Logistic	**50.6250%**	50.6250%	50.6250%	**51.2145%**
SMO	50.6250%	50.6250%	50.6250%	49.8834%
J48	50.6250%	50.6250%	50.6250%	49.4172%
RF	50.6250%	50.6250%	50.6250%	49.4172%

Table A.25: Hotel Reviews To Social Media-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.6250%	**50.6250%**	**50.6250%**	50.3800%
Logistic	**50.6250%**	50.6250%	50.6250%	**51.2145%**
SMO	50.6250%	50.6250%	50.6250%	49.8834%
J48	50.6250%	50.6250%	50.6250%	49.4172%
RF	50.6250%	50.6250%	50.6250%	49.4172%

Table A.26: Hotel Reviews To Twitter-16 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.6250%	**50.6250%**	**50.6250%**	50.3800%
Logistic	**50.6250%**	50.6250%	50.6250%	**51.2145%**
SMO	50.6250%	50.6250%	50.6250%	49.8834%
J48	50.6250%	50.6250%	50.6250%	49.4172%
RF	50.6250%	50.6250%	50.6250%	49.4172%

Table A.27: Social Media To Blog-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	50.0000%	46.1538%	47.1538%	**51.3810%**
Logistic	**56.1538%**	47.6190%	46.619%	47.6190%
SMO	50.0000%	**51.5385%**	**51.5385%**	49.7676%
J48	50.0000%	43.0769%	43.1769%	50.0612%
RF	50.7692%	50.7823%	50.2783%	51.0204%

Table A.28: Social Media To Hotel Reviews-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**51.7196%**	48.2540%	49.2540%	50.2885%
Logistic	50.5732%	51.2500%	51.2500%	**50.3846%**
SMO	50.6774%	**52.9964%**	**52.1964%**	49.7676%
J48	51.1725%	51.8760%	51.2760%	50.2885%
RF	51.0943%	51.5385%	50.5385%	49.3510%

Table A.29: Social Media To Twitter-16 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**50.6250%**	50.6250%	50.6250%	50.6250%
Logistic	50.6250%	50.6250%	50.6250%	**50.6250%**
SMO	50.6250%	**50.6250%**	**50.6250%**	50.6250%
J48	50.6250%	50.6250%	50.6250%	50.6250%
RF	50.6250%	50.6250%	50.6250%	50.6250%

Table A.30: Twitter To Blogs-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**47.6923%**	47.6923%	47.6923%	47.6923%
Logistic	52.3077%	52.3077%	47.6923%	**47.6923%**
SMO	47.6923%	**47.6923%**	**47.6923%**	47.6923%
J48	47.6923%	47.6923%	47.6923%	47.6923%
RF	47.6923%	47.6923%	47.6923%	47.6923%

Table A.31: Twitter To Hotel Reviews-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**50.4950%**	50.4950%	50.4950%	**50.6971%**
Logistic	52.3317%	50.4950%	50.4950%	50.0240%
SMO	50.4950%	**50.4950%**	**50.4950%**	50.0817%
J48	50.4950%	50.4950%	50.4950%	50.0721%
RF	50.4950%	50.4950%	50.4950%	49.5433%

Table A.32: Twitter To Social Media-14 Gender Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**50.5843%**	50.5843%	50.5843%	49.4191%
Logistic	50.5843%	50.5843%	50.5843%	**50.8004%**
SMO	50.5843%	**50.5843%**	**50.5843%**	49.7676%
J48	50.5843%	50.5843%	50.5843%	49.5482%
RF	50.5843%	50.5843%	50.5843%	49.6514%

A.6 Results for Age Identification

A.7 Age Traditional n-grams of Part-of-Speech Tags (Same Genre)

Table A.33: Blogs-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	**46.2585%**	42.1769%	40.1361%	**42.8571%**	**45.5782%**
Logistic	39.4558%	42.1769%	**42.2653%**	41.1837%	42.9796%
SMO	42.8571%	**46.2585%**	42.1769%	42.1769%	40.1361%
J48	36.7347%	35.3741%	29.2517%	31.9728%	33.3333%
RF	44.2177%	33.3333%	43.5374%	42.8571%	36.0544%

Table A.34: Hotel Reviews-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	23.2732%	24.5728%	24.0674%	23.8508%	25.4874%
Logistic	**26.9314%**	**25.1923%**	24.1298%	23.240%	24.9904%
SMO	25.8484%	24.645%	24.5728%	**24.8375%**	**25.5355%**
J48	23.3694%	21.8532%	22.9844%	24.5487%	24.1396%
RF	22.8159%	27.1239%	**26.5945%**	23.5620%	25.3670%

Table A.35: Social Media-14 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	30.6610%	25.7165%	21.9339%	21.3013%	25.0065%
Logistic	**31.5905%**	29.4717%	30.5492%	28.3098%	29.9424%
SMO	30.7041%	30.1879%	29.1010%	29.7823%	31.0007%
J48	30.7772%	28.7116%	28.1952%	28.1952%	28.2081%
RF	30.1188%	**30.2608%**	**31.0225%**	**30.6610%**	**33.2946%**

Table A.36: Twitter-16 Gender Identification, Results using Tn-grams of POST (Same Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	40.9302%	**41.6279%**	**42.093%**	41.1628%	40.6977%
Logistic	33.9535%	34.9535%	34.0535%	33.9535%	32.9535%
SMO	36.2791%	35.1163%	37.907%	**41.3953%**	40.6977%
J48	35.3488%	35.1163%	32.7907%	34.6512%	37.6744%
RF	**41.3953%**	40.6977%	40.0000%	41.1628%	**40.6977%**

A.8 Age Traditional n-grams of Part-of-Speech Tags (Cross Genre)

Table A.37: Blogs To Hotel Reviews Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	21.7088%	23.9952%	24.1396%	23.9230%	24.0193%
Logistic	21.6125%	**24.9579%**	24.1155%	23.3990%	24.0279%
SMO	**23.9952%**	24.7413%	24.3803%	**24.3321%**	24.3321%
J48	23.4176%	22.5993%	**24.5247%**	23.8026%	24.1155%
RF	23.9711%	24.0193%	24.0433%	24.1396%	**24.2599%**

Table A.38: Blogs To Social Media-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	28.5309 %	24.0382%	13.8394%	27.0849%	12.6001%
Logistic	28.1877%	28.3873%	28.9801%	29.2840%	29.8735%
SMO	28.8794%	29.6798%	30.4673%	**30.2091%**	29.8735%
J48	**29.0473%**	23.9737%	**30.5061%**	28.1694%	27.5110%
RF	27.9112%	**29.7444%**	30.1704%	30.0155%	**30.3253%**

Table A.39: Blogs To Twitter-16 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	28.5309 %	24.0382%	13.8394%	27.0849%	12.6001%
Logistic	28.1877%	28.3873%	28.9801%	29.2840%	29.8735%
SMO	28.8794%	29.6798%	30.4673%	**30.2091%**	29.8735%
J48	**29.0473%**	23.9737%	**30.5061%**	28.1694%	27.5110%
RF	27.9112%	**29.7444%**	30.1704%	30.0155%	**30.3253%**

Table A.40: Hotel Reviews To Blogs-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	21.0884%	**29.3958%**	**35.3741%**	**35.3741%**	28.5714%
Logistic	28.7007%	28.4626 %	30.1020%	27.1429 %	31.0204%
SMO	**33.3333%**	23.6638%	24.4898%	26.5306%	**31.2925%**
J48	27.2109%	27.4218%	28.2993%	33.7415%	29.8980%
RF	27.2109%	27.8912%	25.8503%	31.9728%	28.5714%

Table A.41: Hotel Reviews To Social Media-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	27.5884%	**29.3958%**	**29.1118%**	**29.1118%**	**28.3114%**
Logistic	**27.6659%**	28.5740%	29.1058%	28.1936%	27.7418%
SMO	27.0204%	23.6638%	23.9995%	26.6331%	26.2587%
J48	26.0263%	27.9027%	28.1716%	27.1936%	28.2416%
RF	25.6261%	27.7821%	28.2856%	26.2071%	26.5040%

Table A.42: Hotel Reviews To Twitter-16 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	27.5884%	**29.3958%**	**29.1118%**	**29.1118%**	**28.3114%**
Logistic	**27.6659%**	28.574%	29.1058%	28.1936%	27.7418%
SMO	27.0204%	23.6638%	23.9995%	26.6331%	26.2587%
J48	26.0263%	27.9027%	28.1716%	27.1936%	28.2416%
RF	25.6261%	27.7821%	28.2856%	26.2071%	26.5040%

Table A.43: Social Media To Blogs-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	16.3265%	**36.7347%**	8.1633%	7.4830%	7.4830%
Logistic	23.8095%	19.2298%	25.8503%	19.0476%	23.8095%
SMO	20.4082%	20.4082%	19.4082%	26.5306%	30.2925%
J48	25.2109%	27.4218%	**24.2993%**	**29.2517%**	21.7687%
RF	**26.5306%**	31.2925%	15.3309%	23.8095%	**30.6122%**

Table A.44: Social Media To Hotel Reviews-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	16.2696%	18.9170%	19.2298%	19.2780%	19.2058%
Logistic	17.9543%	19.2298%	23.9711%	**25.2708 %**	**23.1288%**
SMO	17.9543%	23.6638%	24.4898%	24.5306%	22.2925%
J48	20.2109%	**27.4218%**	28.2993%	18.8448%	18.3153%
RF	**21.6606%**	20.6739%	**33.3333%**	15.4513%	15.4031%

Table A.45: Twitter To Blogs-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	**39.4558%**	30.6122%	27.2109%	29.9320%	22.9844%
Logistic	38.2993%	36.4626%	38.3810%	33.8231%	30.7007%
SMO	34.6939%	35.3741%	**40.1361%**	**38.0952%**	31.9728%
J48	30.6122%	39.4558%	35.3741%	34.0136%	**36.0544%**
RF	36.0544%	**40.8163%**	39.4558%	36.0544%	33.3333%

Table A.46: Twitter To Hotel Reviews-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	24.0193%	24.1396%	**24.0674%**	23.1047%	22.9844%
Logistic	22.3317%	21.3702%	21.8942%	22.9327%	22.0240%
SMO	**24.8616%**	23.5379%	**23.4898%**	23.8508%	24.0193%
J48	24.3321%	23.6342%	23.0084%	**24.2359%**	23.9471%
RF	24.4284%	**24.2359%**	23.4176%	23.7304%	**24.0433%**

Table A.47: Twitter To Social Media-14 Age Identification, Results using Tn-grams of POST (Cross Genre)

	N-gram Size				
Classifier	**1**	**2**	**3**	**4**	**5**
NB	26.8913%	25.8456%	28.9956%	26.3362%	22.4490%
Logistic	28.5164%	26.3702%	**29.1996%**	28.6514%	27.8004%
SMO	29.3571%	27.8337%	**29.3700%**	28.7245%	**28.8020%**
J48	27.9499%	26.0263%	29.0602%	**28.8794%**	28.1436%
RF	**29.9380%**	**28.9956%**	29.0473%	28.2081%	**27.0849%**

A.9 Age Syntactic n-grams of Part-of-Speech Tags (Same Genre)

Table A.48: Blogs-14 Age Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	32.3077%	39.2308%	36.1538%	41.5385%
Logistic	32.3077%	33.8462%	41.1837%	42.9796%
SMO	40.0000%	**40.7692%**	44.6154%	**44.6154%**
J48	34.6154%	43.0769%	**44.6154%**	41.5385%
RF	**38.4615%**	46.1538%	44.6154%	36.9231%

Table A.49: Hotel Reviews-14 Age Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	21.5737%	24.6222%	22.8508%	25.4874%
Logistic	25.1923%	24.1298%	23.240%	24.9904%
SMO	**26.2637%**	24.2835%	24.0229%	**25.5355%**
J48	23.7103%	24.2835%	**24.5487%**	24.1396%
RF	24.6222%	**25.6384%**	23.5620%	25.367%

Table A.50: Social Media-14 Age Identification, Results using Sn-grams of POST(Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	28.4861%	**31.8359%**	21.3013%	25.0065%
Logistic	29.4717%	30.5492%	28.3098%	29.9424%
SMO	**29.2132%**	27.7331%	29.7823%	**31.0007%**
J48	27.7331%	28.0966%	28.1952%	28.2081%
RF	31.8359%	31.2646%	**30.6610%**	30.2946%

Table A.51: Twitter-16 Age Identification, Results using Sn-grams of POST (Same Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	27.8125%	32.8125%	34.6875 %	34.6875%
Logistic	29.3750%	34.0535%	33.9535%	32.9535%
SMO	**38.7500%**	**36.5625%**	35.9375%	36.8750%
J48	31.8750%	36.5625%	**36.8750%**	**39.6875%**
RF	33.4375%	32.8125%	34.0625%	31.5625%

Table A.52: Blogs To Hotel Reviews-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	23.9952%	24.1396%	23.923%	24.0193%
Logistic	**24.9579%**	24.1155%	23.399%	24.0279%
SMO	24.7413%	24.3803%	**24.3321%**	**24.3321%**
J48	22.5993%	**24.5247%**	23.8026%	24.1155%
RF	24.0193%	24.0433%	24.1396%	24.2599%

Table A.53: Blogs To Social Media-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	28.5309 %	24.0382%	13.8394%	27.0849%
Logistic	28.1877%	28.3873%	28.9801%	29.2840%
SMO	28.8794%	29.6798%	30.4673%	**30.2091%**
J48	**29.0473%**	23.9737%	**30.5061%**	28.1694%
RF	27.9112%	**29.7444%**	30.1704%	30.0155%

Table A.54: Blogs To Twitter-16 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	34.375%	34.375%	34.375%	34.375%
Logistic	**39.6875%**	**39.6875%**	**39.6875%**	**39.6875%**
SMO	34.3700%	34.375%	34.375%	34.375%
J48	34.3750%	34.375%	**3**4.375%	34.375%
RF	34.375%	34.375%	34.375%	34.375%

Table A.55: Hotel Reviews To Blogs-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	21.0884%	**35.3741%**	**35.3741%**	28.5714%
Logistic	28.7007%	30.102%	27.1429%	31.0204%
SMO	**33.3333%**	24.4898%	26.5306%	**31.2925%**
J48	27.2109%	28.2993%	33.7415%	29.898%
RF	27.2109%	25.8503%	31.9728%	28.5714%

Table A.56: Hotel Reviews To Social Media-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	27.5884%	**29.1118%**	**29.1118%**	**28.3114%**
Logistic	**27.6659%**	29.1058%	28.1936%	27.7418%
SMO	27.0204%	23.9995%	26.6331%	26.2587%
J48	26.0263%	28.1716%	27.1936%	28.2416%
RF	25.6261%	28.2856%	26.2071%	26.504%

Table A.57: Hotel Reviews To Twitter-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	34.375%	34.375%	34.375%	34.375%
Logistic	**39.6875%**	**39.6875%**	**39.6875%**	**39.6875%**
SMO	34.3700%	34.370%	34.37%	34.37%
J48	34.3750%	34.375%	34.375%	34.375%
RF	34.3750%	34.375%	34.375%	34.375%

Table A.58: Social Media To Blogs-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	**36.7347%**	8.1633%	7.483%	7.483%
Logistic	19.2298%	**25.8503%**	19.0476%	23.8095%
SMO	20.4082%	19.4082%	26.5306%	30.2925%
J48	27.4218%	24.2993%	**29.2517%**	21.7687%
RF	31.2925%	15.3309%	23.8095%	**30.6122%**

Table A.59: Social Media To Hotel Reviews-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	18.917%	19.2298%	19.278%	19.2058%
Logistic	19.2298%	23.9711%	**25.2708 %**	**23.1288%**
SMO	23.6638%	24.4898%	24.5306%	22.2925%
J48	**27.4218%**	28.2993%	18.8448%	18.3153%
RF	20.6739%	**33.3333%**	15.4513%	15.4031%

Table A.60: Social Media To Twitter-16 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	39.6875%	39.6875%	39.6875%	39.6875%
Logistic	39.6875%	39.6875%	**39.6875%**	**39.6875%**
SMO	39.6875%	39.6875%	39.6875%	39.6875%
J48	**39.6875%**	39.6875%	39.6875%	39.6875%
RF	39.6875%	**39.6875%**	39.6875%	39.6875%

Table A.61: Twitter To Blogs -14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	36.1538%	36.1538%	36.1538%	36.1538%
Logistic	36.1538%	36.1538%	**36.1538%**	**36.1538%**
SMO	36.1538%	36.1538%	36.1538%	36.1538%
J48	**36.1538%**	36.1538%	36.1538%	36.1538%
RF	36.1538%	**36.1538%**	36.1538%	36.1538%

Table A.62: Twitter To Hotel Reviews-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	24.0193%	**24.0674%**	23.1047%	22.9844%
Logistic	22.3317%	21.8942%	22.9327%	22.024%
SMO	**24.8616%**	23.4898%	23.8508%	24.0193%
J48	24.3321%	23.0084%	**24.2359%**	23.9471%
RF	24.4284%	23.4176%	23.7304%	**24.0433%**

Table A.63: Twitter To Social Media-14 Age Identification, Results using Sn-grams of POST (Cross Genre)

	N-gram Size			
Classifier	**2**	**3**	**4**	**5**
NB	25.8456%	28.9956%	26.3362%	22.449%
Logistic	26.3702%	29.1996%	28.6514%	27.8004%
SMO	27.8337%	**29.3700%**	28.7245%	**28.8020%**
J48	26.0263%	29.0602%	**28.8794%**	28.1436%
RF	**28.9956%**	29.0473%	28.2081%	27.0849%