بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

In the Name of Allāh, the Most Gracious, the Most Merciful

**Recite! In the Name of your Lord (ALLAH), who has created (all that exists), has created man from a clot (a piece of thick coagulated blood). Read! And your Lord is the Most Generous, Who has taught (knowledge) by the pen taught man that which he knew not.**

**Surat Al-'ALAQ' (The Clot), the Holy QURAN.**

# A Study of Content Based Methods for Author Profiling in Multiple Genres

**Thesis Submitted to**

**Superior University, Lahore**

In Partial fulfillment of the

Requirement for the Degree of

## Master of Science in Computer Science

By


## Muhammad Waqas Anjum

**MSCS-F14-018**


**Session: 2014-2016**


## Department of Computer Science & Information Technologies

## Superior University Lahore


## Aug 2017

# A Study of Content Based Methods for Author Profiling in Multiple Genres

By


## Muhammad Waqas Anjum

**MSCS-F14-018**

**Session: 2014-2016**


**Thesis Submitted to**

## Superior University, Lahore

In partial fulfillment of the

Requirements for the degree of

## Master of Science in Computer Science

Approved By:

| | |
|---|---|
| _____ | _____ |
| **Mr. Kashif Mahmood** | **Mr. Waqas Arshad** |
| **Program Manager, Research Degrees** | **Research Supervisor** |
| _____ | _____ |
| **Dr. Muhammad Usman Hashmi** | **External Examiner** |
| **Head of Research Department** | **Prof. Dr. Amjad Farooq** |
| **Superior University, Lahore.** | **Associate Professor UET, Lahore.** |

**Thesis Submitted to**


## Department of Computer Science & Information Technologies

# Superior University Lahore

DECLARATION TO BE FILLED BY THE STUDENT AT THE TIME OF SUBMISSION OF THESIS TO THE SUPERVISOR AND/OR FOR EXTERNAL EVALUATION

| Section 1: Particular of the Student | | |
|---|---|---|
| 1.1 | Full Name | **Muhammad Waqas Anjum** |
| 1.2 | Father's Name | **Muhammad Anjum Iqbal** |
| 1.3 | Roll. Number | **MSCS-F14-018** |
| 1.4 | Program | **Master of Science in Computer Science** |

| Section 2: Particular of the Thesis | | |
|---|---|---|
| 2.1 | Title | **A Study of Content Based Methods for Author Profiling in Multiple Genres** |
| 2.2 | Supervisor's Name | **Mr.Waqas Arshad** |
| 2.3 | Date of Completion | **27-08-2017** |


## Department of Computer Science & Information Technologies

## Superior University Lahore

### SUPERVISOR'S CERTIFICATEON

### THESIS SUBMITTED BY A STUDENT

| Section 1: Particulars of the Supervisor | | |
|---|---|---|
| 1.1 | Full Name | **Mr.Waqas Arshad** |
| 1.2 | Address | **The Department of Computer Science & Information Technologies<br>Superior University Lahore.** |

| Section 2: Particulars of the Student | | |
|---|---|---|
| 2.1 | Full Name | **Muhammad Waqas Anjum** |
| 2.2 | Father's Name | **Muhammad Anjum Iqbal** |
| 2.3 | Roll Number | **MSCS-F14-018** |
| 2.4 | Program | **Master of Science in Computer Science** |

| Section 3: Particulars of the Thesis | | |
|---|---|---|
| 3.1 | Title | **A Study of Content Based Methods for Author Profiling in Multiple Genres** |

| | | |
|---|---|---|
| 3.2 | Date of Completion | **27-08-2017** |

I certify that:

a. The above named student has completed the cited thesis under my guidance and supervision.
b. I am satisfied with quality of the student's research work, and
c. I consider it worthy of submission for external evaluation.

| | | |
|---|---|---|
| 4.1 | Supervisor's Full Signature | |
| 4.2 | Date | **27-08-2017** |

## Declaration of Originality

I Muhammad Waqas Anjum hereby solemnly declare that this thesis:

a) is my original work, except where otherwise acknowledged in the text
b) has not been published earlier and
c) shall not be submitted by me in future for obtaining any degree from this or other university or institution
d) has been incorporated HEC Plagiarism Policy
e) In case of violation of HEC Plagiarism Policy, I shall be liable to punishable action under the plagiarism rules of HEC.

| | | |
|---|---|---|
| 3.1 | Student's Full Name | **Muhammad Waqas Anjum** |

| | | |
|---|---|---|
| 3.2 | Student's Full Signature | |
| 3.3 | Date | **27-08-2017** |

# DEDICATION

I thank to All Mighty **ALLAH** (SWT) who created us and make us able to use abilities and strength to complete this task as a final thesis of MSCS. This dissertation is lovingly dedicated to my beloved and most respected Prophet **MUHAMMAD (P.B.U.H)**, beloved parents, wife, uncle, brother; Respected Teachers, Class fellows, and Friends and all those who supported me in all the circumstances helped me and prayed for me. Their support, encouragement, and constant love have sustained me throughout my life.

# ACKNOWLEDGEMENT

First of all, I am very grateful to **ALLAH Almighty** for blessing me with the health, courage, and knowledge throughout the course of this study. I bow my head in deep gratitude before **ALLAH Almighty** for Blessing me with the wisdom and the capabilities and granting me the strength to accomplish this thesis. I would like to express my special thanks to Mr. Waqas Arshad Lecturer Superior University Lahore, who guided me throughout the course of this study and kept me motivated for in time completion of this research work. Without his guidance, help, motivation, and support, this thesis would not have been completed. I would also like to express my sincere gratitude to Superior University, Lahore for providing me with all the necessary facilities. I am very thankful to my beloved parents, wife, brother; Respected Teachers, Class fellows, and Friends for their continuous support, care, love, and prayers throughout my life and specially during this study.

**Muhammad Waqas Anjum**
MSCS-F14-018

# ABSTRACT

## A Study of Content Based Methods for Author Profiling in Multiple Genres

Author profiling is the task of automated prediction of one or more author traits from his/her text. Author profiling is basically grouping of written document on the base of their similarity, content, topic and semantic tags of their author. Further than the identification and verification of a specific person whose writing style is examined, when we talk about author profiling that means how a specific person interact in a social circle and how they share their language so that group them according to their writing style[1]. Automatically detecting an author's profile from text has potential applications in marketing, forensic analysis and detecting harassment cases. To develop and analyze automatic techniques for author profiling, we need a bench mark corpora. In recent years, standard evaluation resources have been developed for different genres including tweets, blogs, hotel reviews, social media etc. This work describes the methods we have employed to solve the author profiling task on datasets. The proposed system is based on simple content based features to identify an author's age, gender and other personality traits. The problem of author profiling was treated as a supervised machine learning task. First content based features were extracted from the text and then different machine learning algorithms were applied to train the models. Results showed that content based features approach can be very useful in predicting the author's traits from his/her text.

## Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Author Profiling

Author profiling is the process of identification of a person's gender, age, native language, personality traits and other demographic in order from his/her written text[1]. We are living in this era where knowledge is growing so quickly arising many difficult problems for researchers one of the problems is author profiling. Now most of the textbook is online. People write and share their opinions ideas behind the curtain limit of secrecy. The author profiling has become an important problem in the fields like linguistic forensics, marketing and security. The work of author profiling is to classify one or more author traits and an author profile compose of the resulting a set of calculating traits. Significantly to the distinguish of the authors attribution, the author profiling task is to be expected even when the written by the author are not in the training data. In the contrary of author attribution, the greater result can be expected when the training data contains from the written text of the authors, because each trait of the model are then expected to be more vigorous. Automatic finding of an author's profile from his/her text has become an emerging and popular research area in now days. Automatically calculating the characteristics of the authors from their texts has a number of potential applications. Such as forensics analysis (Corney et al., 2002; Abbasi and Chen, 2005), advertising intelligence (Glance et al., 2005) and sentiment analysis.

The traits in Encyclopedia of Psychology point of view is defined as "individual differences in the characteristic patterns of thinking, feeling and behaving." it's judicious that an individual identity is fundamental set of qualities, so individuals are comparative in few characteristics but diverse in others. This range can be reflected through the writer one's speech, writing, images etc. In the authorship analysis deals with the finding out the techniques to identify these differences and use of them to distinguish profile of the author such as gender, age, native language, education, profession or personality type. We aim to apply content based approach on a variety of corpora which are on different genres including social media, hotel review, twitter and blogs.

The content may stand up other than material, how author is feeling while writing it, it may reflect distinctive behavioral examples of its author. The information possessed by the content can be utilized by distinctive machine learning methodologies/algorithms to consequently perceive identity sorts.

Authorship analysis can be of two types:

(1) The authentication of the author's assignment where the approach of the personality for the authors is to be observed, to check wither a writing belong to specific author or not.

(2) Author profiling differentiates among the classes of writers studying their collective feature, that is, how language is shared by the people. This helps is to recognizing profiling features such as gender, age, native language, education, profession or personality type. The author profiling can be defined as given the set of texts you have to identify age, gender, profession, education, native language and similar personality traits.

Author profiling is the matter of interest for many different areas such as psychology, linguistics and natural language processing. 21st century is surely the century of science and technology. Internet and web technology has shortened the distances and open new pathways for learning and sharing information worldwide mainly inform of electronic text and videos. As a result of this advancement in social media we observe a lot of unlabeled content as well, which means though there are heaps of contents available, but we lack information about generator of the content. In this respect, the author profiling is trying to conclude the age, mother tongue, education, work or behavior of the authors by discover their available texts.

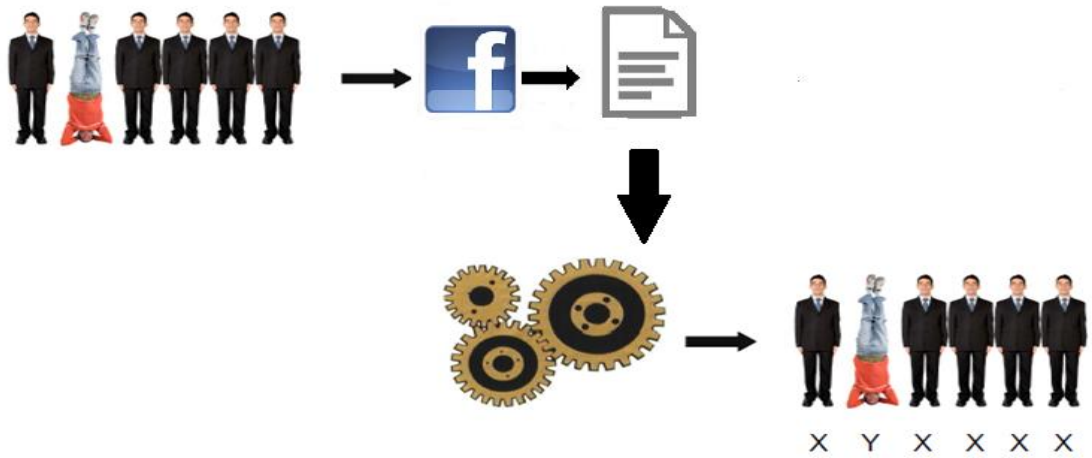



**Figure 1.1 Author Profiling**

## 1.2 Importance of Author Profiling

In today's world the society medium has a great impact on the lives of people across the world. It provides collective networking sites which can help people to connect with their friend's family and acquaintances. This worldwide connectivity provides lot of advantages like seeking new jobs, receiving medical advice by consulting specialists online and free advertising of their products etc. Face book is the one of the most famous social networking sites. Face book was launched on February 4, 2004 and it is serving till today and allowing a large number of clients a chance to share their ideas thoughts and experiences in the shape of comments, posts, pictures, and recordings. According to a study that Face book had 1.59 billion monthly active users in

2015 (www.Statista.com, www.zephoria.com) and now there are over 1.86 billion monthly active Face book users. Although face book has provided us a way to connect with people living in different places in the world but it has a downside too, now we don't have privacy anymore. Now more close chance to the world which may increase the risk of fraud and identity theft by fake accounts. According to an estimation made by Face book that 8.7 percent (83.09 million), of its accounts do not belong to real profiles[2].It is important in many cases to know that who is the author of a given profile. Author profiling can be useful in such cases. For example from Forensic perspective, the purpose of the multilingual person of a man's profile who composed a "doubtful content" can give helpful information about the suspect. The reviews of unknown people about products can help the companies to know about the personality behavior and choices of their customers. From a security perspective, linguistic assessment may profile possible criminals.

## 1.3 Application of Author Profiling

Author profiling can be useful in following fields:

- **Marketing**

In the marketing point of view, the companies may be paying attention to know, on the trends in their products, what kinds of people resembling or dislike their goods, what is necessary by the customer and how to provide and which customer category/class they can attract more. By analyzing the anonymous feedback on product the company want to judge about the type of personality.

- **Forensics**

From the Forensic point of view, the determination of the multilingual profile of a person who wrote an "undoubtfull text" can present in order to useful their previous data.

- **Security**

Multilingual profile may help to overcome the possible criminal activity. Security also related to defense.

## 1.4 Background

In previous times electronic medium was not so much sophisticated that there was no requirement for its analysis. With the passage of time the advancement in technology the use of internet has change the lives of people. Computer mediated communication (CMC) which

includes emails, blogging, chat rooms, conversation boards and social media has presented the huge channels for online communication. So the use of (CMC) like emails, blogs, websites and Social Media has increased, that's why the amount of text is incremented so the need to identify "who is who" has also become important and it's a field of grow up attention these days.

## 1.5 Methods for Author Profiling Detection

There are two methods is used for author profiling detection.

- **Manual Detection** – time consuming and costly.
- **Automatic Detection** – you need training and testing data to develop an automatic system.

## 1.6 Thesis Outline

The remaining part of this thesis organized into the following four chapters which are given below.

**Chapter 2** This chapter we describe an overview of the existing work in the PAN author profiling and also explains the existing bench mark author profiling corpora and also the other author profiling corpora like British National Corpora, Blog corpus for gender prediction . we also provides the previous work in author profiling, Statistics of corpora and also explain the standard author profiling techniques.

**Chapter 3** This chapter will presents the experimental setup, author profiling techniques applied on the corpus, and evaluation measures used to evaluate the performance of methodology used for identifying author personality. This chapter describes the Datasets that are used for PAN author profiling also discussing the Content Based technique and the benefits of the feature selection.

**Chapter 4** This chapter contains the Results and the comparison.

**Chapter 5** Discuss the conclusion and future work.

# Chapter 2

# Literature Review

## 2.1 Existing Author Profiling Corpora

### 2.1.1 PAN 2013 author profiling corpus

There were 21 Participants in PAN2013 competition for author profiling task. The paper describes the corpus and its characteristics. This Paper the author extracts Corpus of the written text from British National corpus. Corpus consists of blogs, tweets which is in the form of text. But in English language; English speakers were describe their experiences or thought more in elaborated way. Different content features were also used by many participants. Different participants considered named entities, emotion words, and sentiment words. In gender prediction the author compares functional words with the POS and achieving the 80% accuracy. Author profiling is increasing significance day by day in a verity of areas such as forensics, marketing and security. This paper the author takes out the corpus from the written text from British National corpus. Corpus consists of blogs, tweets in the form of text. Corpus is a collection of the written texts specially the entire work of the particular author.

The PAN 2013 competition is to determine the age and gender identification of the author from the given document of the data set provided by the PAN. The Author profiling Task at PAN-AP2013, the identification of age and gender based on a very huge amount of corpus collected from the social media.

Author profiling is help out to find out an author's age, gender native language and personality etc. The PAN 2013 has two genres which are blog post and chat logs. There are two different languages used in the PAN2013 which is English and Spanish. The linguistic features vary according to the profile of authors. This Paper the author take out the Corpus of the written text from British National corpus. Corpus consists of blogs, tweets which is in the form of text. Corpus is a collection of large amount of written texts especially the entire work of particular author. Author profiling task at PAN 2013, the identification of age and gender based on a large corpus collected from social media. There were 21 participants take part in the PAN 2013 competition. The data collection about the number of author is in the form of following parts which is Training, Early Bird and Test. Regarding the age group the author make the following classes 10s (13-17), 20s(23-27) ,30s(33-47). Sexual Predators were also discussed in the PAN 2013.

The 21 participants are to identify age and gender of unidentified authors with the help of large and practical collection of blog posts and chat logs. The participants used several different

features or techniques to solve the problem. They used following approaches techniques which are given below. Content Based approach (bag of words, named entities, dictionary words, slang words, contractions, sentiment words, emotional words, and so on), Stylistic based (frequencies, punctuations, POS, HTML, use readability measures and many other statistics), n-grams based, IR based and collocations-based . The author compares functional words with the POS and achieving 80% accuracy in gender prediction. Author profiling is increasing significance day by day in a verity of areas such as forensics, marketing and security.

| Traits | | |
|---|---|---|
| | Hotel Reviews | |
| Age | 10s | 17200 |
| | 20s | 85799 |
| | 30s | 133597 |
| Gender | Male | 118296 |
| | Female | 118300 |

**Figure 2.1 Statistics (PAN 2013 Corpus)**

### 2.1.2 PAN 2014 author profiling corpus

Author profiling is growing importance day by day, such as forensics, security, weather forecast, business etc. There were 10 Participants in PAN 2014 competition for author profiling task. The paper describes the corpus and sub corpus which is divided into four parts. In PAN 2014 the evaluation process is divided into two types one is early birds and second was final evaluation. There were seven submission of early birds and 10 for final evaluation. In PAN 2014 the results were separately given for the evaluation process of each corpus for each language. In this PAN 2014 competition the results were given in accuracy of identification of age and gender. In early bird case the best result were achieved for Twitter. There was no big difference in results of both English and Spanish language. The concerned with blogs there were similar results for gender identification. But in case of age the best results were obtained in Spanish language. The results as concerned of social media, the Spanish results were better than for all the predications of English language. The highest accuracies in gender identification got Spanish social media, Twitter texts and hotel reviews. In gender and age identification the highest values were

achieved by shrestha14 on Spanish Twitter with 0.8846 and in age identification are 0.6923, 0.6154 in joint identification of both age and gender.

The best results in the final evaluation were achieved for Twitter in early bird. For gender identification accuracies are greater in English language while age and joint identification are higher in Spanish language. All the results are lower than the early birds results. The best results in gender identification were achieved in English language and the best results for age identification were achieved in Spanish language. The lowest results obtained in Spanish blogs for age identification. The marquardt14 who obtained worse results but all other participants achieved better results in English.

The result of hotel reviews and social media are very close to the result of early bird. The highest values were achieved by liau14 in gender identification on English Twitter (accuracy of 0.7338) and by shrestha14 in age identification on Spanish Twitter (accuracy of 0.6111) as well as in joint identification on Spanish Twitter (accuracy of 0.4333).The technique bag-of-words used by liau14. The team shrestha14 used word-grams technique. The villenaroman14 achieved the best results in all genres used by the term vector. The marquardt14 used the content and style features and gave the good results in gender identification in Spanish Twitter and in Spanish Blogs. In general prospective for author profiling the use of n-grams technique is not to be good approach. The team lopezmonroy14 used the second order representation and shows overall best performance. Twitter texts achieved the good results because of high number of Tweets as compared to other genres. There are lowest result of the hotel reviews and social media because of lowest accuracy rate of age and gender identification. The comparison of PAN 2013 and PAN 2014 were not possible because of different age group. Some of the approaches could not be work properly and the results were compiled by the only executed approaches. The PAN 2014 competition were better from the previous PAN 2013 competition. The highest and better results achieved in Twitter due to the larger number of Tweets and lowest result achieved in hotel reviews and social media because of lowest results of age and gender identification.

| 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 |
| Total files | | 4160 | 147 | 7746 |

**Figure 2.2 Statistics (PAN 2014 Corpus)**

### 2.1.3 PAN 2015 author profiling corpus

In PAN 2015 there were 22 participants submitted software and 21 teams submitted the notebook papers. In PAN 2015 the corpus from Twitter calculated in four languages Spanish, English, Dutch, and Italian. The corpus is explaining in the age classes' gender and personality traits only in English and Spanish only. The age and gender information were also given the Twitter user. In this year of PAN 2015 competition the dataset were divided into three categories, early birds, for Training and for test. The Dutch dataset were the highest accuracies with the 90% in most of the cases. The Dutch dataset contains the smallest number of authors. The lowest performance contains the English dataset which were the greater in number. The team alvarezcarmona15 performance was better in all the four language. This team used the approach of second order representation (SOR). This team also performs quite well in PAN 2013 and PAN 2014. The team gonzalesgallardo15and grivas15 was also getting the results very close to the team alvarezcarmona15. These three teams was the top three in all the languages. The team of gonzalesgallardo15 uses the two techniques char and POS, n-grams. The second team used the technique of TF-IDF and n-grams, also used style based feature.

In gender identification the greatest value achieved by the alvarezcarmona15 (0.9659), and on second highest values achieved by the grivas15 (0.9432) and on third position achieved by the kiprov15 (0.9091) which is the extra ordinary. The results of Italian and Dutch were similar for the age and personality traits. In this year of PAN competition there was obtained higher accuracies for the both age and gender compared to the shared task of the previous years of PAN competitions. The reason behind this the number of Tweets is enough to calculate the age and gender. Regarding the personality traits the best results with 95% obtained in Italian and Dutch languages.

| 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 |
| **Total files** | | **152** | **34** | **100** | **38** |

**Figure 2.3 Statistics (PAN 2015 Corpus)**

### 2.1.4 PAN 2016 Author Profiling Corpus

In PAN 2016 the main purpose to achieve the prospective of the age and gender identification of the cross genre. The training data is given from the twitter and also from the different corpora like social media, essays, and blogs and in this competition the 22 participants were evaluated. Firstly built the trained model on one genre of the twitter and then evaluated with the different genre of the twitter. There were 22 participants has to identifying the age and gender cross genre in the English, Dutch, and Spanish languages. These were used the following features apply to find out the problem which are following Stylistic and content based feature. Some of them used the second order representation which gives the best results in the first three edition of the PAN. In the final evaluation the Dutch team got the first position in the gender identification. In some of the task the team used the second order representation achieved the first position in some of the assignments.

## 2.2 Other Author Profiling Corpora

### 2.2.1 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 is from the years 1960-1974.

### 2.2.2 Blogs Corpus for Gender Prediction

With the continues development in the social media, the attention shifted to other kind of writings, more informal, less organized and structured, just like blogs. The datasets, on which different experiments were made, contains 566 documents from British National Corpus.

### 2.2.3 Existing Corpora (state of the art)

| Corpus | Language | Traits | Doc Type | Collection methods | No. of Author profiles |
|---|---|---|---|---|---|
| **PAN-13** | English, Spanish | Age, Gender | Blogs , Hotel reviews | Free blogs | 71000 blogs |
| **British National Corpus** | British English | Gender | New collection | National and regional news paper | Above 100 million word text corpus |
| **Tweets based Corpus** | English | Age, Gender | Blogs, emails, twitter | bloggers.com | 3250 blogs |

**Table 2.1 Existing Corpus**

## 2.3 Previous work in Author profiling

The majority of the existing work in author profiling has been done mostly based on the bases of a small number of author traits, for example gender and age [3], gender [4], extraversion and neuroticism [5], extraversion, geniality, neuroticism, and conscientiousness [6], extraversion, agreeableness, conscientiousness, neuroticism, and openness. The author describes a testing in judgment the author's gender in email with machine learner model, namely Support Vector Machine (SVM). Overall, their approach differentiates the male and female authors satisfactorily. The main result of that functional word the most important evidences for the differentiating the gender. They tried to differentiate extraversion from introversion and high neuroticism from low neuroticism in familiar texts. The four sets of features (modality, conjunctive phrases, lexical features, and appraisal) and Support Vector Machine for the classification. The authors tried to predict age and gender in the blog data. Due to the very big in number of authors in the corpus that is greater in the number (18,000), a traditional regular simple classification techniques that was not more feasible. So researchers agreed to for an information retrieval approach using various expression frequency-inverse, document frequency weights, with the combination of this for resemblance. The technique was not able to predict the author correctly in approximately70% of the attempted predictions. The problem when investigating the author profiling problem is too necessary to get labeled data for the authors, to obtain their age and gender. In the classical literature deals with a very small number of recognized authors, the labeling can easily be applied by manually, for this aspect the social media data this is a more complicated task.

The authorship profiling problem is growing importance day by day specially in the few years .There are many author profiling applications such as in security ,commercially and in forensics. In the police department the author profiling play a key role to avoid crime. In market point of view or the large companies the author profiling technique knowing about the like or dislike the product of that company. This survey is based on the online reviews or blogs [5]. Stylistic based features and content based features are the two most basic types of features. There are many other features are also included in the textual type of style like lexical, vocabulary complexity and syntactic, based features whereas in some special case there are also following features are also used such as grammatical, morphological, and orthogonal errors in the simple text.[5]

There are many techniques involve in the computer forensics investigation including data mining, hiding analysis, link and casual analysis, and timeline correlation and many more. E-mail plays a vital role to communicate with the employers or persons in the departments/organizations. There have many departments companies and government officials have written communication within the organization or outside the company. E-mail used for many purposes such as message information swap the documents and also used for legal actions [4]. E-mail can be evidence in many cases such as fraud, sexual harassment, threats, and so on. E-mail can also be misused and cause the incorporate message or information or to incomplete documents is also distributed from this. Content based profiling on simple usual text analysis method is “bag of words and when they use the other naïve bayes method have more difficult rather than the bag of words. In the author identification for the purpose of the e-mail forensics has limitations due the bad categorization presentation approach because of the large amount of data has to be needed for the particular author or not used the good decision to distinguish the other author [4].

Many recommender systems are used to the method of collaborative filtering to the different user’s performance. Content based approach used the information and makes the suggestions. The advantage of this approach is to untreated items of the user to the single interest and provides the details for its recommender. In this paper the content based book recommending the author used the machine learning algorithms for information extraction for the text categorization [2]. Recommender system growing importance because of they suggest films, music, books and other so many things that decide the likes or dislike on the basis of recommending approach. There are many online book stores or shopping palaces like Amazons, Noble and many others used the recommendation approach and providing the history of the renders advisory services. Social filtering or collaborative filtering in the existing recommender systems utilizes a form of computerized matchmaking technique. LIBRA is the first content based book recommender used the simple Bayesian learning algorithm to get the extraction of the books information from the web. Initially this approach is good to produce the perfect recommendations. Finally the machine learning research method developed the ideas that are especially useful for the content based recommending filtering and text categorization associated as well as collaborating approach [2].

Author profiling is the job of knowing the gender, age, native language, or the personality of the author that is how they behave socially. The approach of Machine learning is determining the unknown the age, gender, and personality of the author. In the PAN 2013 the researcher predicting the age and gender from the blogs from applying the different types of the features such Content Based, topic based and Style based and after applying these features the accuracy rate for the age is 64.08% and 64.30%. The accuracy rate for the gender is 56.53% and 64.73% of the English and Spanish respectively [6]. The text from the user's profile is getting importance and helps the many fields like marketing and forensics. The marketing point of view the company manager easily find out the age and gender of different people that like or dislike their products and getting the view point of the public [6]. So in this paper the author trying out the relation of the author's and their language style used by the consumer. The key idea behind all this exercise to analysis how these languages effects in the socially point of view. In this paper the author build the model by using the writing style and content written blogs. A good system is required in the many of the domains for author profiling to get better results accuracy.

The major aim of author profiling is to discover the most relevant information of the person by his/her written text [7]. There are the following categories that analysis strictly how to identify-ying the author's age, gender, language, education and his/ her behavior in social and economically. There are many features that can calculate the text of the author but the absolute length of the text find out the simple features that are given below:

- Number of Words
- Number of Characters
- Number of Sentences

In author profiling automatically calculating from the written text has a lot of applications [8]. For example in the marketing intelligence point of view the author profiling plays a vital role to sale out the company products by the use of their survey [8]. In author profiling the text preprocessing consist of two parts first is segmentation and other is punctuation study [8]. The e-mail firstly split into paragraph and then paragraph further split into tokens and sentences.

Author profiling is the challenging task and now growing importance day by day and applying in several applications such as marketing, forensics, terrorism prevention, security and also used for the unknown suspicious text etc [9]. The prediction of age, gender, and the personality traits published by the author of their tweets has described in the PAN 2015. Their tweet available in the following languages which have English, Spanish, Italian and last one is in Dutch languages [9]. The syntactic n-gram technique gives the good results on tweets of author but cannot be more successful about the age and gender of the author. For the better improvements regarding the accuracy of the regarding the syntactic n-gram the following approach should be applies. Firstly the size of the tweets should be less or divided into parts and in this we can handle the imbalance the data. Secondly the grammatical mistakes should be handled rather than ignoring them. Thirdly the comparison of the tweets has to combine with the syntactic n-grams and to

merge with the proposed features of the other features that have distinct nature like semantic and lexical features.

In PAN 2016 the main purpose to achieve the prospective of the age and gender of the cross genre [10]. The training data is given from the twitter and also from the different corpora like social media, essays, and blogs and in this competition the 22 participants were evaluated. Firstly built the trained model on one genre of the twitter and then evaluated with the different genre of the twitter. There were 22 participants has to identifying the age and gender cross genre in the English, Dutch, and Spanish languages. These were used the following features apply to find out the problem which are following Stylistic and content based feature. Some of them used the second order representation which gives the best results in the first three edition of the PAN. In the final evaluation the Dutch team got the first position in the gender identification. In some of the task the team used the second order representation achieved the first position in some of the assignments.

## 2.4 Statistics of corpora:

| Articles name /(Corpus) | Doc Type | Language | Traits | Collection Methods | No. of author profiles | Techniques Applied |
|---|---|---|---|---|---|---|
| **TAT (training corpus)** | Emails | Arabic | Age, Gender, Education, Extraversion, Lie, Neuroticism, Psycoticism . | English emails Including native and non-native speakers. | 1030 from Arabic speakers | Morphological analysis, lexicon taggers, entity recognition. |
| **TAT (training corpus)** | Emails | English | gender, age, geographic origin, education level, native lang. | English emails Including native and non-native speakers. | emails from 1033 English speakers | Morphological analysis, lexicon taggers, entity recognition. |
| **Chat Mining for Gender Prediction** | chat server (Heaven BBS), | Turkish | Gender Prediction | Chat messages (typed in Turkish) | 250,000 chat messages | Content based, style based |
| **Effects of Age and Gender (PAN-13)** | Blogs, hotel reviews | English , Spanish | Age, Gender | all blogs accessible from blogger.com | 71,000 blogs. | Content based, style based, IR, n-grams |
| **Improving Gender Classification of Blog Authors** | Free blogs | English | Gender | blogger.com, technorati.com | 3100 blogs. | |

**Figure 2.4 Statistics of corpora and techniques applied**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Tweets based corpus** | Blogs, twitter. emails | English | Age, Gender | blogger.com, technorati.com, | 3226 blogs | |
| **Using syntactic features to predict author personality from text** | Essay | English | gender, mother tongue or region, | Easy written by students and demographic info from MBIT test | 145 BA student essays | |
| **Large Scale Personality Classification of Bloggers** | Free blogs | English | neuroticism , extraversion , openness to experience , agreeableness conscientiousness | Bloggers data from different sites. | large collection of around 3000 bloggers | |
| **Classifying author personality from weblog text** | personal weblog | English | Neuroticism, Extraversion, Openness, Agreeableness Conscientiousness | Bloggers data from different sites. | 71 participants (47 females, 24 males | |

**Figure 2.5 Statistics of corpora and techniques applied**

## 2.5 Standard author profiling techniques

There are three most popular author profiling techniques were used in the most recent previous work and we also used these three techniques are:

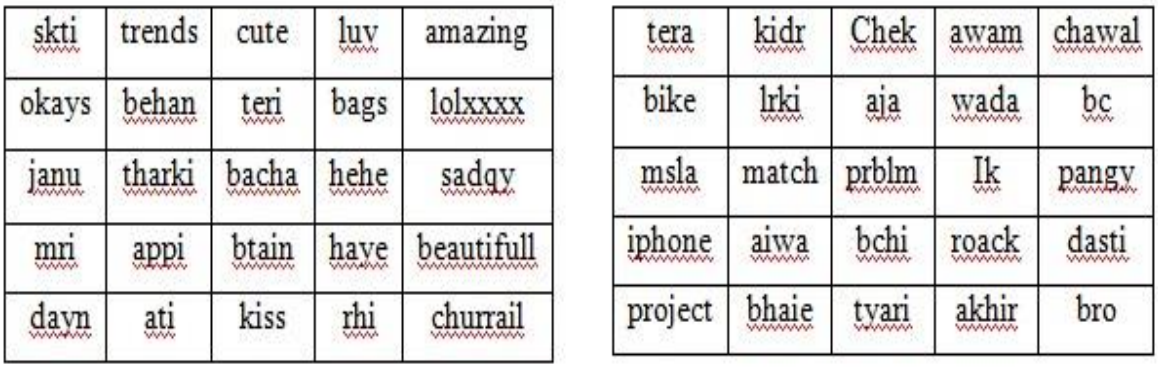
- Stylistic Based approach
- Content Based approach
- Topic Based approach

### 2.5.1 Stylistic Based Approach:

Stylistic-based technique is a very well-known method for authorship attribution problem. This technique has been used by many researchers who have worked on the topic of author profiling. In the past for identifying different demographic traits used for a writer. stylistic based features contain frequencies, punctuation, parts of speech, HTML, read ability measures and many different statistics[2]. It also contains function words i.e. syntactic features or complexity based features, the function words with parts-of-speech (POS) when get combined has proved to be quite more successful (Argamon et al., 2003; Koppel et al., 2002).

## 2.5.2 Content Based Approach

The alternative of two opposite genders like male and female always have different opinion and this difference can be seen in their written contents too. Male authors mostly like to talk about politics, news while female authors are more interested in fashion, shopping, dressing and parties. The features of the content based technique can be useful to differentiate the male and female authors (Schler et al., 2006). For example, a text that contains content related to cricket is more likely to be written by a male author rather than a female author. In a text amount of words that are used most frequently like Imran khan, cricket, IPL, BMW, football, Pakistan etc. Therefore a very huge number of words like these will increase the probability to more chances of it has been written by male author rather than female author. In the same way the huge number of words or phrases like my mehndi, bridal shower, mother in law, satrangi lawn sale, khaddi, stylo high heels etc will chance the probability of it has been written by female author. Similarly if most of the words are like my kids, shopping mall, nail paint etc. This word shows that there is a great chance that it can be written by the female author. The words which are used commonly by one group when contrasted with other can be used as features. It has been seen that youngsters are more interested to talk about topics like video games, school, those who lies in 20s age group like to speak about college life ,Movies and individuals of 30s age group mostly like to write about their occupation, marriage life and politics.



| skti | trends | cute | luv | amazing |
|---|---|---|---|---|
| okays | behan | teri | bags | lolxxxx |
| janu | tharki | bacha | hehe | sadqy |
| mri | appi | btain | have | beautifull |
| dayn | ati | kiss | rhi | churrail |

| tera | kidr | Chek | awam | chawal |
|---|---|---|---|---|
| bike | lrki | aja | wada | bc |
| msla | match | prblm | Ik | pangy |
| iphone | aiwa | bchi | roack | dasti |
| project | bhaie | tyari | akhir | bro |

**Figure 2.6 Content based Features (Author Profiling)**

Along these lines content based components are imperative to distinguish writings having a place with various age groups. Content based technique include components which has used of the following like bag of words, words n-grams, term vectors, named substances, lexicon words, slang words, constrictions and conclusion words.

### 2.5.3 Topic Based Approach

Topic based approach is to write about different topics as well as to express themselves differently about the same topic. For topic based approaches, first the topics from a document are extracted (for example using MALLET toolkit). From all those three techniques, we implemented content based technique on our corpus as they are more famous of the three.

## 2.6 Summary Table

| Content Based | Stylometry Based | Topic Based |
|---|---|---|
| It considers Content of the text. | It considers the writing style of the text | It chooses words from text and analysis topic of the text. |
| Each word is a feature. | It has around 64 unique feature based on writing style. | All key words consider as features. |
| It becomes ambiguous when all authors are writing on same topic. | It becomes ambiguous while comparing current writing and writing of 10 year back of same author. | It is ambiguous when multiple topics are discussed in same topic. |

**Table 2.2** Summary table to compare existing techniques

## 2.7 Weka Classifiers:

- J48
- Random Forest
- SMO
- Naïve Bayes
- ID3

## 2.8 Accuracy

Preposition of correctly identifying instances is called accuracy. Accuracy is the percentage of true outcomes (true positives and two true negative) among the total number of cases examined. To make clear the semantic context, it is often referred to as "rand precision" .It is a test parameter.

accuracy=

$$\frac{numberoftruepositives+numberoftruenegatives}{numberoftruepositives+numberoftruenegatives+numberoffalsepositives+numberoffalsenegatives}$$

Accuracy= (TP+TN) / (TP+TN+FP+FN)

Here,

TP is number of true positives

TN is number of true negatives

FP is number of false positives

FN is number of false negatives

The performance of the author profiling solution for age and gender will be ranked by accuracy.

Proportion of correctly identifying instances is called accuracy. In our research we use accuracy as an evaluation measure for calculating the results of features.

# Chapter 3

## Datasets & Methodology

### 3.1 Introduction

In this chapter presents the experimental setup (corpus used for experimentation), author profiling techniques applied on the corpus, evaluation measures used to evaluate the performance and methodology used for identifying author personality. In this chapter also shows the results and analysis of applied approaches.

### 3.2 Datasets

The corpora which have been explained above in the chapter 2 detailed for running the experiments. Statistics of corpora that I have selected are explained below:

### 3.2.1 PAN 2013 author profiling corpus

The PAN 2013 dataset consists of total 236596 files. The PAN 2013 contains only the blogs data in different age groups (10s, 20s, and 30s) with gender male and female.

### 3.2.2 PAN 2014 author profiling corpus

Blogs, social media, twitter and hotel reviews are the four different geners used in PAN14.The PAN 2014 dataset consists of total 12053 files in three different categories Hotel reviews, Blogs and Social media. The PAN 2014 Hotel reviews corpus is 4160 files, PAN 2014 corpus for Blogs is 147 files and PAN 2014 corpus for Social Media corpus is 4476 files of total dataset.

### 3.2.3 PAN 2015 author profiling corpus

The PAN 2015 dataset consists of twitter tweets in four different languages English, Dutch, Spanish and Italian. English corpus for author profiling task consists of 152 files, PAN 2015 Dutch corpus for author profiling task consists of 34 files, In PAN 2015 Italian corpus for author profiling task consists of 38 files and PAN 2015 Spanish corpus for author profiling task consists of 100 files.

## 3.3 Comparison of Author Profiling Techniques used by different Authors in PAN

| | Second Order Representation | Stylistic and Content Features | N-Grams | Linguistic Features | POS Features | Information Retrieval | Readability Features | Unsupervised Features | Emoticons | Content based Features | Word Gram | Bag of words |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PAN 2013** | The team Meina used the second order representation but the team De-Arteaga got the first position in gender and team Pastor L. perform well in age. | There are so many participants that are used the Stylistic and content feature. | Four Participants used N-Gram technique. | This approach used only one team with the collaboration of the other team. | Five Participants used this approach. | There is one team used this approach | Commonly used in several approaches. | _____ | Only used two participants and discarded by one participants | This features used many participants. Most of the team used these techniques but they did not get better results. | ____ | ____ |
| **PAN 2014** | The team lopezmonroy14 used the second order representation | marquardt14 used the content and style features | One team used N-gram technique | _____ | Use this feature | One participants used this approach | _____ | __ | _____ | _____ | The team shresth a14 used word-grams | The technique bag-of-words used by liau14. |
| **PAN 2015** | The winner of this year competition uses this technique. The team alvarezcarmona15 used the second order representation and got the first position. | Many participants use this approach with the combination of n-gram models. | The ten teams used this approach. | _____ | Different participants used this features. | Use this feature | _____ | ___ | Use this feature | The authors used topic Modeling with Latent Semant-ic Analysis (LDA). | ____ | Use this feature |

**Table 3.1 Comparison of AP techniques used by different authors in PAN**

In PAN 2013 the Team De-Arteaga shows the good results in English and Spanish task because they used SOR. This seems to be good in English language and this feature do not seems to be good for the Spanish language.

- The majority of approaches regarding Stylistic and content features were used and get the results and placed differently in the grade.
- The content based features were also used by so many teams that participate in the PAN-AP13.
- For unsupervised features used to introduce a new variety to build the Corpus for the first time used in PAN-AP13.
- Readability features used for the majority of the teams that are participants and achieved the eight positions in English task and the thirteenth in Spanish task in PAN-AP13.
- N-gram techniques did not achieve the good results.
- There is the number of participants who identified the correct cases of adult-adult sexual discussion.
- In this PAN 2014 the results were given in correctness rate of identification of age and gender of English task.
- In early bird case the best result were achieved for Twitter. There were no big differences in results of both English and Spanish language.
- The comparison of PAN 2013 and PAN 2014 were not possible because of different age group.
- The team lopezmonroy14 used the second order representation and shows overall best performance.
- In general prospective for author profiling the use of n-grams technique is very bad approach.
- Twitter texts achieved the good results because of high number of Tweets as compared to other genres. The lowest result of hotel reviews and social media because of lowest results of age and gender identification.
- The technique bag-of-words used by liau14.
- The team shrestha14 used word-grams technique.
- The villenaroman14 achieved the best results in all genres used by the term vector.
- The comparison between the PAN-AP13 and the PAN-AP14 are not feasible due to the age groups that are shows on the author profiling task.
- Dutch datasets were the highest accuracies with the 90% in most of the cases.
- Lowest performance contains English dataset which was the greater in number. The team alvarezcarmona15 performance was better in all the four language. This team used the approach of second order representation (SOR).
- This team also performs quite well in PAN 2013 and PAN 2014. The team gonzalesgallardo15and grivas15 was also getting the results very close to the team alvarezcarmona15. These three teams was the top three in all the languages.

- The team of gonzalesgallardo15 uses the two techniques char and POS, n-grams. The second team used the technique of TF-IDF and n-grams, also used style based feature.
- In gender identification the greatest value achieved by the alvarezcarmona15, and on second highest values achieved by the grivas15 and on third position achieved by the kiprov15 which is the extra ordinary.
- The results of Italian and Dutch were similar for the age and personality traits.
- In this year of PAN competition there was obtained higher accuracies for the both age and gender compared to the shared task of the previous years of PAN competitions. The reason behind this the number of Tweets is enough to calculate the age and gender. Regarding the personality traits the best results with 95% obtained in Italian and Dutch languages.

## 3.4 Evaluation Measure

Accuracy and root mean square error are used for evaluation of results. The normal evaluation measure used for author profiling tasks in recent research is accuracy (Rangel et al., 2013, 2014), so I have selected it as evaluation measure for our experiments as well.

## 3.5 Evaluation methodology

The purpose of these experimentations is to identify the various personality traits of author. For this purpose the problem is treated as supervised classification task. Two types of classification are applied: (I) binary classification task (II) and multi-classification task. In binary classification task, the aim is to distinguish between gender of author: (1) male and (2) female, and to check true or false for traits like: open, agreeable, conscientious, extroverted and stable, While multi-classification task aims to differentiate between age groups which vary in each corpora like for 2014 and 2015 age group are: (1)18-24 (2) 25-34 (3) 35-49 and (4) 50-XX.

For evaluation following steps are followed:

### 3.5.1 Pre-process

For the experiment point of view the data must be clean and all the raw data should be clean and all the ambiguities should be removed. Before using the data for the running experiments data is pre-processed Unicode, XML and HTML tags are removed to extract text. For content based approach a lot of junk words, raw data are removed and different filters are also applied to select subset of best attributes.

### 3.5.2 Features

The techniques mostly used for the automatic identification of an author's personality traits from the text can be categorized into three broad categories: (1) Stylometry based approaches (which aim to identify an author's traits from his writing style), (2) Content based approaches (which identify author traits using features extracted from the content of the document) and (3) Topic based approaches (which try to predict an author's profile based on the topics used in the document). The system which we used in this thesis is based on the content based approach.

## 3.6 Content Based Features

Content based features were selected to specific gender according to age group and calculated the frequencies of some selected unigrams used by participants in author profiles. The approach used for author identification is "content based", for this simply convert given text into word tokens and count the number of their occurrences.

To convert text into words use the "string to word vector" filter available in WEKA toolkit, then customized this filter and compared the results by setting different attributes and applying on a corpus to find best modification for this filter, and then select the best settings for testing on rest of data. There are different content based features like Chi square, information gain, gain ratio are used.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Their | Them | The | that | Terms | that's | there's | themselves |
| Therefore | These | They | tips | Follow | today | too | tools |
| Top | As | Available | Angeles | Analyst | Bank | Big | Biscuits |
| Bites | Black | Country | Click | Colin | Credit | College | Class |
| Costa | Davis | December | Devon | Equality | Face book | Find | Florida |
| All | Air | After | a | As | As | Time | Google |
| Although | American | And | Another | Do | Don't | by | how |

**Table 3.2 Sample of Content Based features used**

## 3.7 Benefits of feature selection:

There are three main benefits of feature selection which are given below.

- **Reduces Over fitting:** Reduces over fitting means less redundant data less opportunity to make decisions based on noise.
- **Improves Accuracy:** Accuracy means a smaller amount to displaying precision misleading statistics increases.

- **Reduces Training Time:** Reduce time means fewer data than the algorithms train quicker.

## 3.8 Feature selection using WEKA

WEKA is software that provides built in algorithms to select attributes; in WEKA just to select the attribute evaluator and then apply search method that want to use.

## 3.9 WEKA toolkit

In this research work used the WEKA toolkit. WEKA is the software for the popular machine learning approach that is written in Java language which contains a visualization collection tools and data analysis algorithms and predictive modelling with graphical user interfaces for easy access to this feature.

WEKA is free software available under the GNU General Public License. WEKA contains tools for:

- data pre-processing
- classification
- regression
- clustering
- association rules
- visualization

### 3.9.1 Classifiers

Classifiers are used to match an input document with a feature value. In this thesis, use machine learning algorithms 5 from the toolbox of WEKA J48, Naive Bayes, Random Forests, SMO support vector machine), and ID3.

### 3.9.2 WEKA Classifiers:

- J48
- Random Forest
- SMO
- Naïve Bayes
- ID3

# Chapter 4

# Results and Analysis

## 4.1 Results and analysis

This section will report and analyze the end results that we have achieved in this section, for our experiments using the content based techniques and also the comparison of the PAN author profiling task.

## 4.2 Content Based Technique

For content based features, we applied the five classifiers; and the results are reported in the table. Overall, the results for content based features are very good and encouraging. The second approach used for author identification is "content based", for this we simply convert given text into word tokens and count the number of their occurrences.

To convert text into words we use "string to word vector" filter available in WEKA toolkit, we customized this filter and compared the results by setting different attributes and applying on a small corpus to find best customization for this filter, and then select the best settings for testing on rest of data.

| **PAN-AP13** | | | |
|---|---|---|---|
| **Sr. No.** | **Classifier** | **Gender** | **Age** |
| 1. | **J48** | 52.3235 % | 48.3098% |
| 2. | **Naïve Bayes** | 54.2083 % | 47.1423 % |
| 3. | **ID3** | 51.9039 % | 49.4391% |
| 4. | **Random Forest** | 55.8794 % | **61.6482 %** |
| 5. | **SMO** | **56.894%** | 51.456% |

**Table 4.1 Results for datasets PAN-AP13**

The table 4.1 shows the results for the datasets of the PAN13. This table shows the results of applying classifiers that is used to for the experiment of Content based approach. The classifier Support Vector Machine (SMO) performs the best result in gender identification with the accuracy of the 56.894%. The classifier Random Forest performs best with the accuracy rate of the 61.6482%. The Naïve Bayes classifier perform badly in age identification with its lower rate 47.14% and the ID3 classifier perform badly in gender identification with its lower rate 51.9039%. The classifiers SMO performs well and ID3 performs badly in PAN13.

| **PAN-AP14** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Blogs** | | **Social Media** | | **Twitter** | |
| Sr. No. | **Classifier** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** |
| 1. | **J48** | 55.102% | 89.795 % | 50.374% | 35.456% | 49.0196% | 42.4837% |
| 2. | **Naïve Bayes** | 61.224 % | 43.537 % | 51.097% | 40.765% | 49.0196% | 42.4837% |
| 3. | **ID3** | 62.585% | 93.8776% | 51.897% | 30.4415% | 49.0196% | 42.4837% |
| 4. | **Random Forest** | 69.387% | **51.23%** | **51.923%** | **51.897%** | **49.6732%** | **42.4837%** |
| 5. | **SMO** | **70.748%** | 51.707% | 49.742% | 49.098% | 49.0196% | 42.4837% |

### 4.2 Results for datasets PAN-AP 2014

The table 4.2 describes the results for the datasets of the PAN14. Blogs, social media, twitter and hotel reviews are the four different geners used in PAN14.The table shows the results of applying classifiers that are used to for the experiment of Content based approach. The classifier Support Vector Machine (SMO) performs the best result in gender identification with the accuracy rate of the 70.748%. The classifier Random Forest performs best with the accuracy rate of the 95.283% in the age identification. The J48 classifier perform badly in gender identification with its lower rate 55.102% and the Naïve Bayes classifier perform badly in age identification with its lower rate 43.537%. The classifiers Random Forest performs well and J48 performs badly in PAN14 with respect to blogs genre.

| PAN-AP15 | | | |
|---|---|---|---|
| Sr. No. | **Classifier** | **Gender** | **Age** |
| 1. | **J48** | 65.5629 % | 54.9669 % |
| 2. | **Naïve Bayes** | 66.8874 % | 68.8742 % |
| 3. | **ID3** | 69.5364 % | 45.6954 % |
| 4. | **Random Forest** | 70.1987 % | 62.2517 % |
| 5. | **SMO** | **76.1589 %** | **73.5099 %** |

### 4.3 Results for datasets PAN-AP15

The table 4.3 describes the results for the datasets of the PAN15. The PAN15 is about gender, age, and the personality traits of the twitter using user's. The table shows the results of applying classifiers that are used to identifying the age and gender for the experiment of Content based approach. The classifier Support Vector Machine (SMO) performs the best result in gender and age identification with the accuracy rate of the 76.1589%, 73.5099% respectively. The classifiers Random Forest performs well and J48 performs badly in PAN14 with respect to blogs genre. The PAN15 were in four different languages Italian, English, Spanish, and the Dutch. The classifier Random forest is in the second position of the PAN15 in gender and the classifier Naïve Bayes is on second 70.1987%, 68.8742% respectively.

| PAN-AP16 | | | |
|---|---|---|---|
| Sr. No. | **Classifier** | **Gender** | **Age** |
| 1. | **J48** | 49.5413 % | 41.7431 % |
| 2. | **Naïve Bayes** | 49.5413 % | 41.7431 % |
| 3. | **ID3** | 49.5413 % | 41.7431 % |
| 4. | **Random Forest** | **49.7706 %** | **41.7431 %** |
| 5. | **SMO** | 49.5413 % | 41.7431 % |

### 4.4 Results for datasets PAN16

The table 4.4 describes the results for the datasets of the PAN16. In PAN 2016 the main purpose to achieve the prospective of the age and gender identification of the cross genre. The training data is given from the twitter and also from the different corpora like social media, essays, and

blogs and in this competition the 22 participants were evaluated. The table shows the results of applying classifiers that are used to identifying the age and gender for the experiment of Content based approach. The classifier Random Forest performs the best result in gender and in the age identification all the other classifier in which J48, Random Forest, ID3, SMO, and Naïve Bayes perform very well according to the situation and performs well. The classifiers Random Forest performs well with the accuracy rate of the 49.7706%and all other classifiers perform also very well with the accuracy rate of 41.7431% with respect to gender and age.

- **4. 3 Comparison with the PAN Datasets:**

| **Our Best Results** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **PAN-AP13** | | **PAN-AP14** | | | | | | **PAN-AP15** | | **PAN-AP16** | |
| | | **Blog** | | **Social Media** | | **Twitter** | | | | | |
| **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** |
| **56.9%** | **61.7%** | **70.8%** | **51.3%** | **51.92%** | **51.90%** | **49.7%** | **42.5%** | **76.1%** | **73.6%** | **49.8%** | **41.8%** |
| **Best results of PAN** | | | | | | | | | | | |
| **PAN-AP13** | | **PAN-AP14** | | | | | | **PAN-AP15** | | **PAN-AP16** | |
| | | **Blog** | | **Social Media** | | **Twitter** | | | | | |
| **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** | **Gender** | **Age** |
| **59.2%** | **65.7%** | **67.9%** | **46.1%** | **54.2%** | **36.5%** | **73.3%** | **50.6%** | **85.9%** | **83.8%** | **75.6%** | **58.9%** |

**Table 4.5 Comparison with the PAN Datasets**

With respect of the English task, the results of the PAN-AP13 our accuracy rate is below in gender than the teams who are participate in the task of author profiling. The difference between our results with the other author profiling participants is just only 3%. The technique which we used to obtain the result is content based technique. Our results are better from the 16 teams out of the 22 participants of the PAN. The results are better from more than 72% of the participant's team. The highest accuracy rate of the PAN-AP teams is 59.2% and our result is 56.7% which seems not to the bad results. We are closer to those results. We were obtaining 72% accuracy for the gender identification. The accuracy rate of the age identification is 61.7% and the accuracy rate of the participant's team is 65.72% as shown in table 4.6.The difference of the results is only by 4% from the participant's team of the PAN-AP13. But our results shows the 68% accuracy rate shows for the age identification. According to these results shown, whether we participate in the PAN-AP13 and get the $4^{th}$ or $5^{th}$ position in the competition. Our result is better from the 15 teams out of 22 participant's teams. The classifier which we used SMO performs well in gender identification 56.9% and the classifier Random Forest performs well in the age identification 61.75 in our research. For all datasets content based techniques are applied for age and gender prediction. The approach used for author identification is "content based", for this we simply convert given text into word tokens and count the number of their occurrences.

To convert text into words we use "string to word vector" filter available in WEKA toolkit, we customized this filter and compared the results by setting different attributes and applying on a small corpus to find best customization for this filter, and then select the best settings for testing on rest of data.

The table 4.6 describes the results of the participant's team of the author profiling PAN-AP14 and compare with our results that are shown above. The table shows the results of applying classifiers that are used to for the experiment of Content based approach. The four different sub corpora namely blogs, social media, twitter, and hotel reviews. The results of blog in terms of the English task gender are shown in the table 4.6. Our technique is performs well in the accuracy of the blog and achieved the best results of all the participants of the PAN-AP14. The accuracy rate of the gender is terms of blogs are 70.8% and the accuracy rates of the teams which participate in the PAN-AP14 are 67.95%. The comparison of our results with that result shows a very good performance in the gender identification in terms of the blog. These results are much better than the other participants and stood in the first rank. The classifier SMO performs well and obtains the accuracy rate 70.8%. The results of the blog in terms of the English task age are shown in the table 4.6. Our results are better from all the team whose participates in the age and gender competition of the PAN. The accuracy rates of the age in terms of social media are 51.3% and the accuracy rate of the teams which participate in the PAN-AP14 are 54.21%. The analysis between these two, our results and the PAN results are close with the difference of the 3%. The classifier we used for this datasets is Random Forest. Our results stood $3^{rd}$ out of 10 teams for the gender identification.

The result of the social media in terms of the English task that our results shown badly in gender identification and perform good in age identification. The result of gender identification is 51.9% and the teams of author profiling result are 54.21%. Our results showed badly performance and only 30% accuracy rate of gender. But in age identification our results are better than the gender identification. The result of age is 51.9% and the teams of author profiling results are 36.52%. The results showed very good performance and achieve the 100% accuracy in age identification.

Our results are better from all the participants of the PAN. The result of the twitter in terms of the English task performs very badly. Our techniques perform very badly in age and gender identification. The result of gender is 49.7% and the teams of author profiling results are 73.38% which is much greater accuracy. The result of age is 42.5% and the teams of author profiling shows good results with the accuracy rate of the 50.65%. Only 50% results of the age identification will be improved. Our results of the PAN-AP14 twitter in terms of accuracy on ENGLISH gender are from the last out of the 10 participants of the PAN and our results in terms of accuracy on ENGLISH age are on $5^{th}$ out of 10 teams of the PAN-AP14.

The table 4.6 describes the results of the participant's team of the author profiling PAN-AP15 and compare with our results that are shown above. The table shows the results of applying classifiers that are used to for the experiment of Content based approach. The results of the PAN-AP15 are not much good. The techniques which we performed perform badly. Both the age and gender identification result are not matched with the team results of the PAN-AP15. Only 59% results in gender identification and 63% results in age identification. Our results are better from the 16 teams out of the 22 participants of the PAN as shown in table 4.6. The accuracy rate of the gender identification is 76.1% and the accuracy rate of the participant's team is 85.92%. Our results are better from the 13 teams out of 22 participants of the PAN. The accuracy rate of the age identification is 73.6% and the accuracy rate of the participant's team is 83.80%.

The table 4.6 describes the results of the participant's team of the author profiling PAN-AP16 and compare with our results that are shown above. The table shows the results of applying classifiers that are used to for the experiment of Content based approach. The results of the PAN-AP16 are not much good. The techniques which we performed the datasets of the PAN-AP16 perform badly. Both the age and gender identification result are not matched with the team results of the PAN-AP16. The result of gender is 49.8% and the teams of author profiling results are 75.64% which is much greater accuracy. The result of age is 41.8% and the teams of author profiling shows good results with the accuracy rate of the 58.97%. Only 10% results will be improved in gender identification and 50% results will improved in age identification. Over all achieved the 60% accuracy from the team participates in the PAN competition.

## 4.4 Summary

In this chapter, we aimed to express the results setup that how the classifiers can be used for the development and analysis of the author profile detection of the PAN. We also explained the comparison of the PAN author profiling tasks in the tabular form and the Comparison of PAN author profiling techniques used by different authors in PAN.

# Chapter 5

# Conclusion and Future Work

## 5. Conclusion

In our research work discussed the Author Profiling technique. This section have covered the role of the main approach that is Content based features in identification of author personality traits. We simply convert text into word vectors and count frequency of each word. We concluded that for content based approach and obtained the best results. The second approach the Stylistic based features are also discussed only for the introduction.

We are quite satisfied about our overall accuracy rate for both age and gender prediction, and hope to improve it further.

## 5.1 Future work

- Sentiment Analysis can be performed on the results of stylistic based and content based features.
- Explore further new techniques that could enhance the work and make it more intelligent for the purpose of future usage makes it more reliable.
- Increase the number of author profiles in corpus.
- Other personality traits can also be explored and detected from text other then age and gender.
- For author profiling using different approaches, future research can focus to investigate more features like, native language, and level of education, living city and many more attributes. Exploring other techniques for author profiling can be another avenue.

# Bibliography

[1] Mooney, R. J., & Roy, L. (2000, June). Content-based book recommending using learning for text categorization. In Proceedings of the fifth ACM conference on Digital libraries (pp. 195-204). ACM.

[2] Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In AAAI spring symposium: Computational approaches to analyzing weblogs (Vol. 6, pp. 199-205).

[3] De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. ACM Sigmod Record, 30(4), 55-64.

[4] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

[5] Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: predicting age and gender from blogs—Notebook for PAN at CLEF 2013. In In Forner et.

[6] Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., & Wives, L. K. (2014). Examining Multiple Features for Author Profiling. JIDM, 5(3), 266-279.

[7] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). Author profiling for English emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07) (pp. 263-272).

[8] Posadas-Durán, J., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., & Pichardo-Lagunas, O. (2015). Syntactic n-grams as features for the author profiling task. Working Notes Papers of the CLEF.

[9] Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. Working Notes Papers of the CLEF.

[10] López-Monroy, A. P., y Gómez, M. M., Jair-Escalante, H., & nor Pineda, L. V. (2014). Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. Cappellato et al.[6].

[11] Patra, B. G., Banerjee, S., Das, D., Saikh, T., & Bandyopadhyay, S. (2013). Automatic Author Profiling Based on Linguistic and Stylistic Features Notebook for PAN at CLEF 2013.

[12] Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

[13] Rangel, F., Rosso, P., Koppel, M. M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at PAN 2013. In CLEF Conference on Multilingual and Multimodal Information Access Evaluation (pp. 352-365). CELCT.

[14] Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daeleman, W. (2014). Overview of the 2nd author profiling task at pan 2014. In CEUR Workshop Proceedings (Vol. 1180, pp. 898-927). CEUR Workshop Proceedings.

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

[16] http://pan.webis.de/ Last visit 23-06-2017

[17] J. Marquardt, G. Farnadi, G. Vasudevan, S. Davalos, A. Teredesai, and M. De Cock, "Age and Gender Identification in Social Media," in CEUR Workshop, 2014.

[18] Yatam, S. S., & Reddy, T. R. (2014, December). Author profiling: Predicting gender and age from blogs, reviews & social media. In International Journal of Engineering Research and Technology (Vol. 3, No. 12 (December-2014)). IJERT.

[19] Najib, F., Cheema, W. A., & Nawab, R. M. A. (2015). Author's Traits Prediction on Twitter Data using Content Based Approach. In CLEF (Working Notes).

[20] Ashraf, S., Iqbal, H. R., & Nawab, R. M. A. Cross-genre author profile prediction using stylometry-based approach. Balog et al.[5].

[21] Rangel, F., Rosso, F. C. P., Potthast, M., Stein, B., & Daelemans, W. Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CEUR-WS. org (Sep 2015), http://www. clef-initiative. eu/publication/working-notes.

[22] Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. Icwsm, 11(1), 281-288.

[23] Juola, P., & Stamatatos, E. (2013, September). Overview of the Author Identification Task at PAN 2013. In CLEF (Working Notes).

[24] Rangel, F. (2013). Author profile in social media: Identifying information about gender, age, emotions and beyond. In Proceedings of the 5th BCS IRSG Symposium on Future Directions in Information Access (pp. 58-60).

[25] Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF.

[26] Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., & Stein, B. (2017). Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017).

[27] Yatam, S. S., & Reddy, T. R. Author profiling: Predicting gender and age from blogs, reviews & social media. International Journal of Engineering Research & Technology (IJERT), ISSN, 2278-0181.

[28] Ciobanu, A. M., Zampieri, M., Malmasi, S., & Dinu, L. P. (2017). Including dialects and language varieties in author profiling. arXiv preprint arXiv:1707.00621.

[29] Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. Literary and linguistic computing, 22(3):251–270, 2007.

[30] Francisco Rangel and Paolo Rosso. Use of language and author profiling: Identifi- cation of gender and age. Natural Language Processing and Cognitive Science, 177, 2013.

[31] Wee-Yong Lim, Jonathan Goh, and Vrizlynn LL Thing. Content-centric age and gender profiling. Proceedings of the Notebook for PAN at CLEF, 2013.

[32] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: Predicting age and gender from blogs. Notebook for PAN at CLEF, pages 119–124, 2013.