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

Muhammad Waqas Anjum Ch, Waqas Arshad Cheema

**Abstract**— Author Profiling is a task of automated prediction of one or more author traits from his/her text. Author profiling is grouping of written document based on 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 interacts 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 harrassment cases. To develop and analyze automatic techniques for author profiling, we need a benchmark corpora. In recent years, standard evaluation resources have been developed for different genres including tweets, blogs, hotel reviews, social media etc.

**Index Terms**— Author Profiling; Authorship; Content Based; Corpus; English; Dutch; Italian.

## 1 Introduction

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 text is online. People write and share their opinions ideas behind the certain limit of secrecy. The problem of author profiling has become an important problem in the fields like linguistic forensics, marketing and security. . Such as forensics analysis (Corney et al., 2002; Abbasi and Chen, 2005), advertising intelligence (Glance et al., 2005) and sentiment analysis. 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. Author profiling is the matter of interest for many different areas such as psychology, linguistics and natural language processing. 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. 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. 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 written by the author is 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.

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.

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 whether 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.

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 opened 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.

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 friends, 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, as 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.

- ***Muhammad Waqas Anjum Chaudhry,*** *is with Computer Science & Information Technologies department as research fellow at* The *Superior University Lahore, Pakistan. He is currently working as faculty member in NCBA&E Lahore, Sub-Campus, Rahim Yar Khan, Pakistan.*
  *(E-mail: waqasch.065@gmail.com).*
- ***Waqas Arshad Cheema*** *is workingas a Faculty member at the Computer Science & Information Technologies department, The Superior University Lahore, Pakistan.*
  *(E-mail: waqas.arshad@superior.edu.pk).*

## 2 RELATED WORK

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 words the most important evidence 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. Because 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 approximately seventy percent of the attempted prediction. The authorship profiling problem is getting importance day by day specially in the few years .There are many authors profiling applications such as in security ,commercially and in forensics. As well in the police department, the author profiling plays 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. 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 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. 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 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 incomplete documents are also distributed from this. Content based profiling on simple usual text analysis method is "bag of words" and when the uses 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]. 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 analyzes strictly how to identifying the author's age, gender, language, education and his/ her behavior in socially and economically. Many features 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 many 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 consists of two parts first is segmentation and other is punctuation study [8]. The e-mail firstly splits into paragraph and then paragraphs 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]. There were 21 participants in PAN2013 competition for author profiling task. The paper describes the corpus and its characteristics. In this paper the author extracts Corpus of the written text from British National corpus. Corpus consists of blogs, tweets that is in the form of text. However, in English language; English speakers described their experiences or thought more in elaborated way. Many participants also used different content features. Different participants considered named entities, emotion words, and sentiment words. In gender prediction, the author compares functional words with the POS achieving the 80% accuracy. Author profiling is increasing significance day by day in a verity of areas such as forensics, marketing and security. The author takes out the corpus from the written text from British National corpus. In this study 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. In PAN 2014, the evaluation process is divided into two types one is early birds and second is final evaluation. There were 7 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 PAN 2014 competition, the results were given in accuracy of identification of the age and gender. In early bird case the best results were achieved for Twitter. The prediction of age, gender, and the personality traits published by the author of their tweets have described in the PAN 2015. Their tweet is available in the following languages, which have English, Spanish, Italian and last one is in Dutch languages [9]. 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 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. The 22 participants have to identify the age, gender and cross genre in the English, Dutch, and Spanish languages. 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. 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. There are three most popular author profiling techniques which used in the most recent previous work but we also used the Content based technique. 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. 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.

| Content Based | Stylometry Based | Topic Based |
|---|---|---|
| It considers content of the text. | It considers writing style of the text. | It chooses words from text and analyses 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 1** Summary table to compare existing techniques

## 3 Data Sets and Methodology

This section represents 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. 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. Blogs, social media, twitter and hotel reviews are the four different genres 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 datasets. In PAN 2015, the English corpus for author profiling task consists of 152 files. The 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. Content-based features were used to select 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 were selected 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 |
| Therefore | These | They | tips | Follow | today |
| Face book | Angeles | Available | As | Analyst | Bank |
| Bites | Black | Country | Click | Colin | Credit |
| Costa | Davis | College | Devon | Equality | Top |
| All | Air | After | a | As | As |
| Although | American | Another | and | Do | Don't |
| there's | cricket | too | tools | Big | Biscuits |
| December | Class | Find | Florida | Time | Google |

**Table 2** Sample of Content Based features used.

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. 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 modeling with graphical user interfaces for easy access to this feature.

## 4 RESULT AND ANALYSIS

This section will report and analyze the results that we have achieved for our experiments using the content-based techniques and the comparison of the PAN author-profiling task. For content-based features, we applied the five classifiers; naming J48, Naïve Byes, Random Forest, SMO, ID3 and the results are reported in the table. Overall, the results for content-based features are very promising 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 selected the best settings for testing on rest of data. With respect of the English language task, the results of the PAN-AP13 and our accuracy rate is below in gender than the teams who have participated 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 obtained 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. The difference of the results is only by 4% from the participant's team of the PAN-AP13. However, our results show the 68% accuracy rate for the age identification. According to these results shown, whether we participate in the PAN-AP13 and get the 4th or fifth position in the competition. Our result is better from the 15 teams out of 22 participant's teams. The classifier that 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. All datasets of the 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. In the PAN-AP14 there are four different sub corpora namely blogs, social media, twitter, and hotel reviews.

| **Our Best Results** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **PAN14** | | | | | | | | | | | |
| **PAN13** | | **Blog** | | **Twitter** | | **Social Media** | | **PAN15** | | **PAN16** | | | |
| **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | | |
| 56.9% | 61.7% | 70.8% | 51.3% | 49.7% | 42.5% | 51.92% | 51.90% | 76.1% | 73% | 49.8% | 41.8% | | |
| **Best Results of PAN-AP** | | | | | | | | | | | | | |
| | | **PAN14** | | | | | | | | | | | |
| **PAN13** | | **Blog** | | **Twitter** | | **Social Media** | | **PAN15** | | **PAN16** | | | |
| **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | **G** | **A** | | |
| 59 % | 65.7% | 67.9% | 46.1% | 73.3% | 50.6% | 54.2% | 36.5% | 85.9% | 83.8% | 75.6% | 58.9% | | |

**Table 3** Comparison with the PAN Datasets

Our technique performs efficiently 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 in terms of blogs is 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. Our results are better from all the team who 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 3rd out of 10 teams for the gender identification. The result of the social media in terms of the English task that our results show 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. However, 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%. 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 fifth out of 10 teams of the PAN-AP14.

The results of the participant's team of the author profiling PAN-AP15 and compares with our results that are shown above in table 3. 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. 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 results of the participant's team of the author profiling PAN-AP16 and compares with our results that are shown above in table 3. The table shows the results of applying classifiers that used to for the experiment of Content based approach. The results of the PAN-AP16 are not much good. The techniques performed the datasets of the PAN-AP16 perform badly. Both the age and gender identification results 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 60% accuracy from the team participates in the PAN competition was achieved.

## 5 CONCLUSION AND FUTURE WORK

In this research paper, we discussed the Author Profiling techniques. This section has 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 followed the 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. It explores further new techniques that could enhance the work and make it more intelligent for the purpose of future usage makes it more reliable in increase the number of author profiles in corpus. Other personality traits can also be explored and detected from text other than 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.

## REFERENCES

[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 Forneret.

[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 02-04-2017

[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).