International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Author Profiling using Stylistic and N-Gram Features

Radha D, Chandra Sekhar P

Abstract: ***The World Wide Web is increasing tremendously with massive amount of textual content primarily through social media sites. Most of the users are not interested to upload their genuine details along with textual content to these sites. To identify the correct information of the authors the researchers started a new research area named as Authorship Analysis. The authorship Analysis is used to find the details of the authors by examining their text. Authorship Profiling is one type of Authorship Analysis, which is used to detect the demographic characteristics like Age, Gender, Location, Educational Background, Nativity Language and Personality Traits of the authors by examining writing skills in their written text. Stylometry is one research area defines a set of stylometric features namely word based, character based, syntactic, structural and content based features for differentiating the author's writing styles. In this work, the experimentation conducted with various stylistic features, N-grams and content based features for gender prediction. These features are used for representing the vectors of documents. The classification algorithms produce the model by processing these vectors. Two classification algorithms namely Random Forest, Naïve Bayes Multinomial were used for classification. We concentrated on prediction of Gender from 2019 Pan Competition Twitter dataset. Our approach obtained best accuracies when compared with many Authorship Profiling approaches.***

Keywords : Authorship Analysis, Authorship Profiling, Accuracy, Content based Features, Gender Prediction, N-grams, Stylistic Features.

I. INTRODUCTION

In the last 20 years, Internet has evolved from a network of connected computers used to share data among researchers. As a result of this growth and the birth of social networks, blogs and many other websites where users are given the opportunity of easily creating or uploading content and the amount of data generated every day has also grown immensely. Most of the generated data in the net is thus unstructured. One of the characteristics of the Internet nowadays is that a user can post anonymously in forums, comment sections of articles, social networks, chat systems, etc. The Authorship Analysis is one research area concentrated by the many researchers to find the details of the authors by analyzing their written textual content.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Radha D*, Department of CSE, Malla Reddy College of Engineering and Technology, Hyderabad, India. Email: radharavavarapu@gmail.com
Chandra Sekhar P, Department of CSE, GITAM, Visakhapatnam, India. Email: chandrasekhar.pothala@gitam.edu

Authorship Analysis is categorized into three techniques such as Plagiarism Detection, Authorship Identification and Authorship Profiling [1]. The Plagiarism Detection detects the percentage of authors contribution is copied from other author's contributions [2]. Authorship Identification classified into two classes namely Authorship Verification and Authorship Attribution. Authorship Verification verifies whether the anonymous document was written by the suspected author or not by investigating the suspected author's documents [3]. Authorship Attribution detects the author of an unknown document by investigating the documents of given set of authors [4]. Authorship Profiling discover the demographic characteristics of an author by investigating the writing style in their texts [5]. In Authorship Identification, the training data need suspected authors documents to recognize the document's author. But in Authorship Profiling, the suspected author's documents need not required in training data to detect the characteristics of the suspected author. This is the major difference among Authorship Identification and Authorship Profiling [7].

Authorship Profiling is used in information processing applications such as harassing messages, forensic analysis, security, educational domain, literary research and marketing [6]. In social websites, people are involved in different crimes like public embarrassment by sending harassing messages, blackmailing, defamation, stalking and creation of profiles with fake details. All these crimes are in the form of messages. The authorship profiling is used here to analyze the harassing messages and detect the basic characteristics like gender, age group, location of messages of authors. In forensic analysis, the forensic experts analyze the property wills and suicide notes to detect the details of the suspected author. In this context Authorship Profiling is one such technique helpful for this purpose. The terrorist organizations send letters and mails to threaten the government bodies. The Authorship Profiling approaches were used in security to analyze these mails and whether the messages came from suspected sources or not. In marketing point of view, the market people are analyzing their products based on the reviews of their product. Based on the analysis of reviews they will take strategic decisions about their products. Authorship Profiling is used to analyze the reviews of products and find the details of reviewers like gender, age, location etc. Authorship Profiling is used in educational domain also. In educational domain, the researchers are able to find the exceptional talented students, the knowledge level of student by analyzing the written texts of the students. In the case of literary and historic studies, Authorship Profiling can be applied to confirm/refute the author characteristics of a text.

Different researchers identified several variances in writing style of female and male by analyzing their datasets. Koppel et al. analyzed [6] large corpus of documents and concluded that woman used more number of pronouns and men used more determiners and quantifiers in their writings. In general, the author style of writing depends on the grammar rules they followed, topics they selected and words they used. They observed that, the female discussed the topics like kitty parties, shopping and beauty in their writings whereas the male concentrated on topics like technology, politics and sports in their writings. In another observation [3], the males more written about politics and technology and female discussed about marriage styles. They also observed that, the adverbs and adjectives usage is more in female writings than male writings.

The words selected for writing a review or a blog is different based on the concept discussed in those sites. There is a possibility of classify a document whether the document was written by female or male based on the words occurred in a document. The presence of words like boyfriend, my husband and pink in a document improves the chances of document written by female authors and the presence of words like cricket and world cup in a text improves the chances of text written by male authors. They also observed that when compared with male authors the female authors used less number of prepositions in their writings [1].

In [8], they found that the experimentation with only content based features was more effective to distinguish the writing style of female and male authors and also identified that the accuracy is reduced when experiment with content based features along with other features. Some other researchers [9] identified that male used more prepositions, articles and longer words in their writings whereas female used more pronouns, verbs, negations and words like friends, home and family.

The dataset used for Authorship Profiling experiments contains small sized messages (Twitter Tweets, harassing messages) and lengthy messages (Reviews, Blogs). The length of the document decides the effectiveness of prediction of characteristics of the authors. Based on the dataset used for experiment, the Authorship Profiling approaches are categorized into two types such as instance based and profile based approaches [6]. In instance based approaches, lengthy messages are used as training data. Each message is considered as one experimental unit or instance. These experimental units are converted into a form which is suitable for analysis. In profile based approaches, the small messages combination is used as training data. In this approach, the small messages of one author stored in one big file. For each author, maintain one file and considered these files as experimental units. These experimental units are converted into a form which is helpful for analysis. In this context, the experimental units are vector representation of text. The classification algorithms are used for analysis of these vectors.

This paper structured in 8 sections. The existing work in Authorship Profiling is discussed in section 2. Section 3 presents the characteristics of the dataset. Section 4 describes the Bag Of Features model. The stylistic features listed in section 5. The results of experimentation are presented in section 6. The results discussed in section 7. The section 8 describes conclusions of this work and possible directions also.

II. RELATED WORK

One of the first studies on performing author profiling with machine learning techniques was done by Argamon et al. [10]. The study observed that the correct linguistic features combination were useful to distinguish the characteristics of an author's text. The scientific event PAN competition arranged several author profiling tasks during the last few years, welcoming anyone with interest to participate. The competition considered different aspects like age, gender, personality traits and native language in author profiling task. SVMs was used by the majority of the participants [11, 12], but also random forest [13], logistic regression [11] and many more algorithms was used. Vollenbroek et al. used [14] linear SVM and achieved the highest average accuracy of 52.58 % by extracting various Stylometry features. However, Modaresi et al. also extracted part-of-speech taggings, various Stylometry features and lexical features, but used those features to train a logistic regression classifier. They got average accuracy of 52.47 % and concluded that part-of-speech taggings were not suitable features when testing on domain independence [15].

The most often explored demographic traits in the literature are clearly gender and age. In [1], the authors describe two experiments on the Blog Corpus to identify gender and age. For both experiments, stylistic features such as function words, part-of-speech frequencies, blog words and hyperlinks as well as the 1,000 most relevant unigrams according to the Information Gain metric are extracted. The chosen classifier is the Multi-class Real Winnow, which achieves 80.1% accuracy for gender and 76.1% for age identification. The same dataset is used in [16], where a stylometric analysis of the age and gender of the bloggers is presented. Two main novel features are used such as sentence length variation and non-dictionary words. The motivation for the second group of features is that after analyzing the word usage per gender and age, the authors concluded that teenagers generally use more non-dictionary words than adults. The final feature set contains the average sentence length, the frequencies of 35 content words as well as 52 slang words. There is an improvement in terms of performance compared to previous approaches that use the same data, obtaining 89.18% in gender and 80.32% in age identification with Naïve Bayes as classifier. Both this approach and the previous one are very content dependent, their systems depend mainly on the specific word choices of the authors.

Argamon et al. present [10] four different experiments in which gender ("man" vs "woman"), age (13-17, 23-27, 33-47), native language (Bulgarian, Czech, French, Russian and Spanish), and personality ("neurotic" vs "non-neurotic") of the authors are predicted. To do so, two kinds of features are extracted such as content based and style based features.

The stylistic features are computed by using taxonomies provided by systemic functional linguistics, which describe meaningful distinctions of function words and parts of speech. The content based features consist of the 1,000 words that have the highest Information Gain coefficient in the training set. Gender classification is further explored in [17]. A new blog post corpus is crawled for this work. F-measure, gender preferential features, stylistic features, word classes and factor analysis as well as part-of-speech sequence patterns are used as features. The best reported result is 88.56% accuracy, achieved using a Support Vector Machines classification algorithm after applying the feature selection algorithm.

Sarawgi et al. present [18] a gender and genre identification model that avoids gender bias in topics. As corpus, blog posts and a collection of scientific papers are used. The accuracies of gender identification in the blog dataset are presented both by topic and in average. The mean accuracy of 68.3% with the character-based model in the cross-topic scenario is the best result. In the scientific domain, character-based features and syntactic features perform equally, achieving 76% of accuracy. In another presented experiment, the system is trained on the blog dataset and tested on the scientific dataset. The accuracy in this experiment decreases significantly due to the differences of the texts used in each phase.

Another work that performs several classification tasks using informal blog posts is [19]. In this case, the chosen language is Vietnamese rather than English. 298 features compose the feature set. In [20], the authors use the British National Corpus to distinguish between female and male authors as well as the genre of the text (fiction vs non-fiction). Part-of-speech frequencies and function words are used as features. The learning method is a variant of the exponential gradient algorithm. The system obtains accuracies of 77.3% in genre distinction and 79.5%-82.6% in gender identification, depending on what genre is used to train the classifier. The authors also experiment with feature reduction algorithms for analyzing the efficiency of the system when the number of features was reduced.

In [22], the authors use the Enron email corpus. To perform gender identification on emails, the authors extract five subsets of features. The total number of features is 545. SVM is used to classify, achieving 82.20% accuracy in the best case. An extension of this work is presented in [21], where the same feature set is used in two scenarios. The first one was already presented in [20]. The second one uses the Reuters Corpus. The experiments predict the gender of the author in both scenarios, obtaining the same accuracy in the first case as in their previous work. In the case of the Reuters Corpus, the authors present an accuracy of 76.75% when SVM is used for classification. The authors conclude that the gender prediction of the author of neutral news is a much more challenging problem than in the case of personal emails. Another approach that uses email messages is described in [22]. 689 features are extracted. The feature set includes character-level features, lexical features and structural features.

Burger et al. present [23] a gender identification approach that uses Twitter data. The selected feature set consists of character 1-5grams and word 1-2grams from the content of the tweets and screen name, full name and description of the profile. This generates more than 15 million distinct features, which presents a challenge to most machine learning toolkits. Using the Balanced Winnow algorithm and some code optimizations, 92% of accuracy is achieved when all the previously mentioned features are combined. The system is also tested using only the content of the tweets. This approach performs worse, obtaining 76% of accuracy.

In [24], a cognitive approach based on neurology studies is presented. The goal is to classify the authors of texts by their gender and age. A study on the frequencies of each grammatical category in 6 different sources (namely Wikipedia, newsletters, forums, blogs, Twitter and Facebook) is outlined. Focusing on Facebook data, the same analysis for each gender is also performed, concluding that in Spanish, men use more prepositions than women and on the other hand, women use more pronouns, determiners and interjections. After these remarks, a feature set is presented composed of word-based features such as the words that start with a capital letter, words with all characters capitalized, word length, etc., usage of punctuation marks, frequency of each part of speech, number of emoticons and the usage of emotion words.

III. DATASET CHARACTERISTICS

The dataset used in this work is collected from the 2019 PAN competition task of Bots and Gender profiling [25]. The details of the tweets dataset for gender prediction are depicted in Table I.

TABLE- I: The characteristics of the dataset

CHARACTERISTICS	Male	Female
Number of Documents	1,030	1,030
Number of Tweets	102,842	102,930
Mean Length (mean number of tokens per document)	2,014	2,123
Vocabulary Length (the number of distinct terms per category)	95,323	102,689

In this competition, the main task of the participants is to distinguish the sources of the text whether it came from human or bot. Once the source is identified the next task of participant is identify the gender of the human. In this work, we are interested in predicting the gender of the human. The training data consists of Twitter tweets of 2060 authors. The dataset is balanced in case of gender that means the number of tweets in both male and female profiles is same. Each author document consists of 100 tweets

IV. BAG OF FEATURES MODEL

The Bag of Features model is used in this work for document vectors representation. The Fig. 1 shows the Bag of Features model.

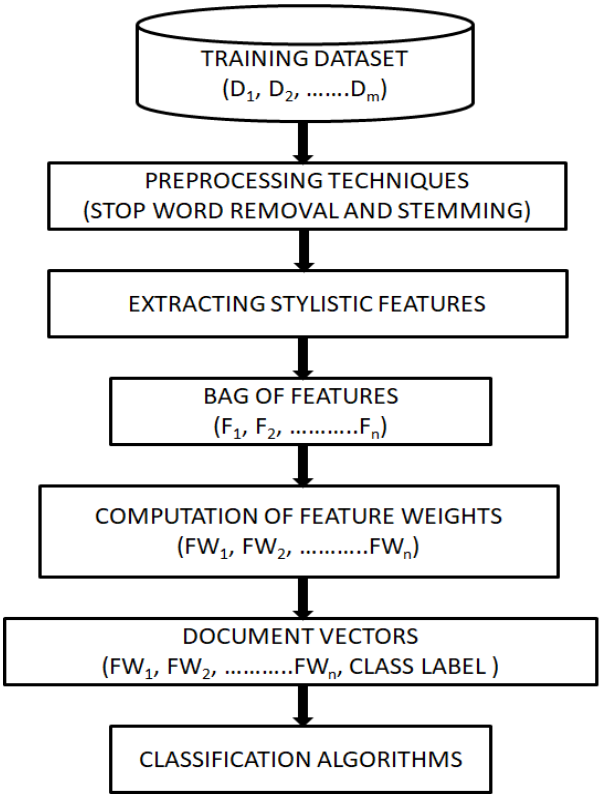

Fig. 1. The Bag of Features Model

In this model, first preprocessing techniques were applied on training data to remove unwanted data or prepare the data for analysis. Two preprocessing techniques namely stop words elimination and stemming was used to clean the dataset. Stop word elimination is removing the words which are not having any discriminative power like the, a, was, between etc. stemming is a technique of converting words into its root. This technique is used to reduce the number of unique words. After preprocessing techniques, extract the suitable stylistic features to distinguish the author's writing style. Consider these features as Bag of Features (F1, F2, ……, Fn). This features set is used for document vector representation. In document vector, each value is denoted with the feature frequency. Finally, the classification algorithms produce the model by training with document vectors and this model used to identify the gender of an author.

V. STYLISTIC FEATURES

To perform Authorship Profiling task, the process of features extraction needs to be carried out. A carefully chosen set of features influence is more to distinguish author's demographic traits and an ineffective set of features shows less influence. The identification of the author of a text or at least characteristics of the author is a very attractive idea because the areas of author profiling have many practical applications. The stylistic features are used to discover the stylistic differences in the textual content of the authors. Various stylistic features such as word based, character based, structural and syntactic features were proposed by the researchers to distinguish the author's writing style. In this work, various stylistic features were used in the experimentation to predict the gender of the authors by analyzing the tweets dataset. It has been shown that the writing style of an author evolves with time, so the writing characteristics of writers are not necessarily static. In this work, four types of features were extracted from tweets dataset for gender prediction.

A. Word Based Features (WBF)

The word based features are based on the words used in a document. Different researchers extracted various word based features in their experimentation. In this work, the experiment performed with the word based features like the counts of words, the positive words, the negative words, unique words count, acronyms count, capital letters words, words starting with capital letters, foreign words count, count of words that occur twice, average word length, maximum length of a word, count of words with numbers, patriotic words ("Indians", "Americans"), words length greater than 6 letters, count of words less than 3 letters, lexical density (percentage of informative words), count of words with repetitive characters, count of contraction words, count of stop words, Type Token ratio, word N-grams (N is 1, 2, 3). The most frequent 200 word N-grams were considered as word based features.

B. Character Based Features (CBF)

The CBF are based on the characters used in the documents. The list of character based features used in this work are characters count, count of capital letters, count of punctuation marks (usage of quotes, periods, commas, colons, semi-colons, ellipsis, question marks, interrogation and exclamation Marks) , digits count, special characters count, Emojis count, character N-grams (N is 2, 3, 4, 5). The most frequent 200 character N-grams were considered as character based features.

C. Structural Features (STF)

The structural features are based on the structure of a document. The structural features used in this work are paragraphs count, sentences count, count of sentences per paragraph, count of words per paragraph, count of characters per paragraph, variation in tweet length, count of HTML tags, count of hashtags, count of Re-tweets, count of @mentions, count of URL's used.

D. Syntactic Features (SYF)

The syntactic features are based on the syntax of the language used in the dataset. The syntactic features used in this work are determiners count, prepositions count, singular nouns count, plural nouns count, proper nouns count, pronouns count, adjectives count, adverbs count, count of spelling errors, count of past verb tenses, count of future verb tenses, Part Of Speech (POS) N-grams (N is 1, 2, 3). The most frequent 200 POS N-grams were considered as syntactic features. There have been many previous approaches that implemented author profiling systems successfully. These approaches have the tendency of focusing on the content of the text instead of on its inner structure. They can be very effective in controlled environments, where every document in the dataset belongs to the same genre and domain is written in the same language and has similar characteristics.

However, content-related approaches do not generalize well and are computationally expensive (using feature vectors of thousands of features). Some researchers believe that a carefully chosen set of features that is able to characterize the author's style of writing was a very effective approach that circumvents some of the shortcomings that previous approaches had. The style of writing reflects the habits of the writers, thinking about themselves and the patterns that writers used to attain their goals. Style includes diction (choice of words), tone, syntax, discourse, punctuation, spelling, voice and many other characteristics.

VI. EXPERIMENTAL RESULTS

Classification is a process of allocating a predefined class label to an unknown document. In this work, two classification algorithms like Random Forest (RF) and Naive Bayes Multinomial (NBM) were used to evaluate the efficiency of gender prediction. WEKA tool is used to implement these classification algorithms. In this algorithms, 10-fold cross validation is used where in the training dataset is randomly divided into 10 samples. In every iteration, 9 samples were used for training the classifier and 1 sample is used to test the performance of the trained classifier. This process is iterated till each sample is acted as test data exactly one time. In this work, Accuracy measure is used to evaluate the classification algorithms. Accuracy is the number of documents predicted correctly their gender in collection of test documents.

In this work, first we extracted suitable stylistic features for distinguishing the male and female style of writing. Each document is represented with these stylistic features as document vectors. The RF and NBM are used for producing the model. The accuracies of the classifiers for gender prediction are presented in table II.

Table- II: The accuracies of gender prediction for stylistic features

Classifier/ Stylistic features	NB M	RF
CBF	66.47	67.13
WBF	70.21	76.87
STF	58.73	61.43
SYF	69.52	73.58
CBF+WBF+STF+SYF	78.64	81.45

The RF classifier performance is good in accuracy of gender when compared with NBM classifier. The CBF, WBF, STF and SYF features individually obtained the accuracy of 67.13%, 76.87%, 61.43% and 73.58% respectively for gender prediction when Random Forest classifier is used. The combination of all these features obtained 81.45% for gender prediction when RF classifier is used.

Next, the experiment continued with content based features. These features depend on the content of a text. In the extraction of content based features, first remove stopwords and perform stemming. After cleaning the text, the dataset contains only informative words. Compute the TFIDF (Term Frequency and Inverse Document Frequency) of each informative word. TFIDF assign more weight to the words which are present in fewer documents. We used top 2000 words based on TFIDF scores in the experimentation. Every tweet document is represented with these 2000 words as a document vector. Two classification algorithms are used to produce the model by training the vectors of documents. The accuracies of the classifiers for gender prediction are presented in table III.

Table- III: The accuracies of gender prediction with TFIDF weighted words

Classifier/ Number of words	NB M	RF
500	64.31	68.45
1000	66.78	71.98
1500	72.46	75.65
2000	75.67	78.32

In table III, the RF classifier achieved 78.32% accuracy for prediction of the gender when most TFIDF scored 2000 words are used as features. The gender accuracy is improved when the number of words is increased. In this experiment also the RF performs better than NBM classifier.

Later, the experiment performed with the N-grams features. N-grams are the overlapping sequence of 'n' number of tokens. Character N-grams, Word N-grams and POS (Part Of Speech) N-grams are the overlapping sequence of 'n' number of characters, words and Part Of Speech (POS) tags respectively. The Part Of Speech (POS) tags are identified by using Stanford POS tagger. We identified 3000 most frequent N-grams (1000 character N-grams, 1000 word N-grams, 1000 POS N-grams) for experimentation. The accuracies of these n-grams for gender prediction are represented in table IV.

Table- IV: The accuracies of gender prediction with most frequent character, word and POS n-grams

Classifier/ Number of features	NBM	RF
Character N-grams (1000)	71.43	73.59
Word N-grams (1000)	76.15	80.87
POS N-grams (1000)	74.71	78.51
3000	81.65	85.33

In table III, the RF classifier obtained 85.33% accuracy for prediction of the gender when most frequent character, word and POS n-grams were considered as features. When experimented N-grams individually, the word N-grams attained best accuracy for predicting gender when compared with character and POS N-grams. The RF classifier obtained good accuracies compared to NBM classifier.

VII. DISCUSSION OF RESULTS

The experiment started with stylistic features of CBF, WBF, STF and SYF. Table II shows the accuracies of stylistic features. The character based features obtained 67.13% accuracy for gender prediction when RF classifier is used. It was observed that the punctuation marks and character N-grams influence the gender prediction accuracy. The word based features obtained an accuracy of 74.87% for gender prediction. The word based features performs well compared
to character based, syntactic and structural features.

The features stopwords count, Type/Token ratio, unique words count, word N-grams influence is high for getting good accuracy for gender prediction. The structural feature performance is less compared to other features used in this work. The syntactic features accuracy is more for gender prediction when compared with structural and character based features. The POS N-grams of syntactic features influence the gender prediction accuracy. The combination of all stylistic features attained an accuracy of 81.45% for predicting gender when experimented with RF classifier.

In this work, 2000 content based features identified based on the TFIDF scores. Table III represents the accuracies of content based features. It was detected that the combination of stylistic features performed well for predicting gender compared with content based features. But the content based features accuracy is more than the individual features accuracies. The experiment continued with the most frequent character, word and POS N-grams of each 1000. The accuracies of N-grams are displayed in Table IV. The word N-grams achieved best gender prediction accuracy when compared with character and POS N-grams. The combination of N-gram features attained best accuracy when compared with the accuracies of content based and stylistic features.

VIII. CONCLUSION AND FUTURE SCOPE

In this work, the experiment started with the various set of stylistic features like word based, character based, structural and syntactic features. The combination of stylistic features attained good gender prediction accuracy. Next, the experiment continued with content based features and detected that the accuracies of content based features are less compared to the accuracies of combination of stylistic features. Finally, the experiment performed with most frequent N-grams of various types such as character, word and POS N-grams. It was identified that the combination of N-grams gender prediction accuracy was good compared with the content based and stylistic features accuracies. In all experiments, the Random Forest classifier performs better than Naïve Bayes Multinomial classifier.

In future work, we concentrated on finding the suitable feature selection algorithm to reduce the features count or find the most informative features. It was also planned to propose a new document vector representation technique to improve the accuracy of gender prediction.

REFERENCES

1. Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. The AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 199–205.
2. Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, Benno Stein: Overview of the 6th International Competition on Plagiarism Detection. CLEF 2014: 845-876.
3. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors., 1261–1276 (Dec 2007).
4. M. Sudheep Elayidom , Chinchu Jose , Anitta Puthussery ,Neenu K Sasi "Text classification for authorship attribution analysis", Advanced Computing: An International Journal, Vol.4, No.5, September 2013.
5. T. Raghunadha Reddy, B.VishnuVardhan, and p.Vijaypal Reddy, "A Survey on Authorship Profiling Techniques", International Journal of Applied Engineering Research, Volume 11, Number 5 (2016), pp 3092-3102.
6. Koppel M. S. Argamon and A. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing, pages 401-412, 2003.
7. E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. JASIST.
8. Pennebaker J. The secret life of pronouns: What our words say about us. Bloomsbury, USA; 2013.
9. Newman ML, Groom CJ, Handelman LD, Pennebaker J. Gender differences in language use. An Analysis of 14,000 Text Samples Discourse Processes; 2008. p. 211–36.
10. Argamon, S., Koppel, M., Pennebaker, J.W., and Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. Commun. ACM, 52(2):119–123.
11. K. Bougiatiotis and A. Krithara, "Author profiling using complementary second order attributes and stylometric features - notebook for pan at clef 2016," in CLEF, 2016.
12. J. D. Rodwan Bakkar Deyab and T. Gon¸calves, "Author profiling using support vector machines - notebook for pan at clef 2016," in CLEF (Notebook Papers/Labs/Workshop), 2016.
13. R. M. A. N. Shaina Ashraf, Hafiz Rizwan Iqbal, "Cross-genre author profile prediction using stylometry-based approach - notebook for pan at clef 2016," in CLEF (Notebook Papers/Labs/Workshop), 2016.
14. T. K. M. M. C. P. J. B. H. H. Mart Busger op Vollenbroek, Talvany Carlotto and M. Nissim, "Gronup: Groningen user profiling - notebook for pan at clef 2016," in CLEF (Notebook Papers/Labs/Workshop), 2016.
15. M. L. Pashutan Modaresi and S. Conrad, "Exploring the e↵ects of crossgenre machine learning for author profiling in pan 2016 - notebook for pan at clef 2016," in CLEF (Notebook Papers/Labs/Workshop), 2016.
16. Goswami, S., Sarkar, S., and Rustagi, M. (2009). Stylometric Analysis of Bloggers' Age and Gender. In ICWSM. The AAAI Press.
17. Mukherjee, A. and Liu, B. (2010). Improving Gender Classification of Blog Authors. The Conference on Empirical Methods in Natural Language Processing, EMNLP '10, pages 207–217, Stroudsburg, PA, USA.
18. Sarawgi, R., Gajulapalli, K., and Choi, Y. (2011). Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 78–86, Stroudsburg, PA, USA.
19. Pham, D. D., Tran, G. B., and Pham, S. B. (2009). Author Profiling for Vietnamese Blogs. In Asian Language Processing, 2009. IALP '09. International Conference on, pages 190–194.
20. Cheng, N., Chen, X., Chandramouli, R., and Subbalakshmi, K. P. (2009). Gender identification from E-mails. In Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on, pages 154–158.
21. Cheng, N., Chandramouli, R., and Subbalakshmi, K. P. (2011). Author Gender Identification from Text. Digit. Investig., 8(1):78–88.
22. Estival, D., Gaustad, T., Hutchinson, B., Pham, S. B., and Radford, W. (2007). Author profiling for English emails. The 10th Conference of the Pacific Association for Computational Linguistics, pages 263–272.
23. Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA.
24. Rangel, F. and Rosso, P. 2013 On the identification of emotions and authors' gender in facebook comments on the basis of their writing style. In CEUR Workshop Proceedings, volume 1096, pages 34–46.
25. https://pan.webis.de/clef19/pan19-web/author-profiling.html

AUTHORS PROFILE

Mrs. D. Radha working as Associate Professor in the Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology. She published research papers in Various International journals and conferences. She has memberships in IAENG and IFERP.

Dr P.Chandra Sekhar is presently working as Associate Professor in Department of CSE, Gitam University . He published several papers in various international conferences and journals. His current research interests are Cryptography, Speech & Image processing.