# A Systematic Exploration of Style Breach Detection Methods during Author Attribution Process

*Rajesh Shardanand Prasad[1], Dr. Midhun Chakkaravarthy[2]*
*[1] Post Doc Fellow, Lincoln University College, Malaysia*
*[2] Principal Supervisor, Lincoln University College, Malaysia*

**Abstract**— Detecting and accrediting authorship of diverse text corpus can be advantageous for numerous tasks and domains comprising bibliometrics, information retrieval and plagiarism detection. Various researchers have illustrated authorship attribution techniques into these domains. Authors here extend the concept by combining authorship information for identifying corpus author to further style breach detection. To do this, researchers review diverse methods from text segmentation to those analyzing the spreading of stylometric features across the document and thereby to predict the number of authors accordingly or style change detection. Some experiments are conducted. This review will create a solid platform that will help implementation of potential applications of stylometry, plagiarism exposure, stylistic fingerprints etc. Authors analyze various research outcomes while highlighting their strengths and weaknesses as a baseline for further study.

**Keywords—** Authorship analysis (AA), Author Profiling, Author Clustering, Cross-domain authorship attribution, Style breach detection**.**

## Introduction

Style Breach Detection task represents an emerging application area that is extensively used in contemporary breach detection systems including email breach detection tools, perimeter breach detection systems, life-lock breach detection systems and community breach detection applications. This process involves, detecting borders where authorship is likely to change inside a document. In contrast to the text segmentation problem that emphasize on locating shifts of topics, the style breach detection systems focus on disclosing boundaries with the help of writing style features. Stylometry is a vital instrument in the field of digital text forensics, particularly in situations where we have anonymous or doubtful text corpus authored by one or more humans. These texts do not have an exterior clue, means or source to demonstrate that which text segment is associated to which author. To put it another way, stylometric methods are used when we may have to discover if the commended authorship of text document truly present in situations but we do not have any outside confirmation possessions.

This paper explores the fundamental task of author identification, before implementing stylometry to create a strong theoretical background. This type of system endeavours to expose the authors of several texts. It is apromising area of research leading to applications in literary research, cyber-security, forensics, and analysis of social media. This survey explores two fundamental issues. First is the unique task of cross-domain authorship attribution, where the texts are of known and unknown authorship related to different domains, and second refers to style change detection, where one-author and multiple-author corpus are to be classified.

In the second task, the follower's literature texts are examined, where a large part of contemporary literature written by casual authors who are inspired by some popular trending works, to enable us control the domain of texts for the first time. A new corpus of follower's literature in English is described first. For the second task, a new data set of questionnaires comprising manifold themes in English is presented. Selected state-of-art techniques are reviewed for the cross-domain authorship attribution task and for the stylealteration detection task. A detail study of useful methods and analytical results are presented in thispaper.

The objective of this survey is encouraged by the mission of correlating fragments of text to their actual authors. In this study, authors emphasize on evaluating the approach humans diverse writing elegances. This investigation will be useful in various arenas including authorship acknowledgment, breach of copyright detection, etc. which practice content features. The emphasis is on the stylometric, i.e. content-agnostic, uniqueness of authors. This section presents insights of related work, which is summarized in following section on next page.



DOI: 10.5373/JARDCS/V12SP6/SP20201026
**Corresponding Author: Rajesh Shardanand Prasad E-mail: rsprasad_viit@yahoo.com*
Article History: Received: Mar 20, 2020, Accepted: June 22, 2020

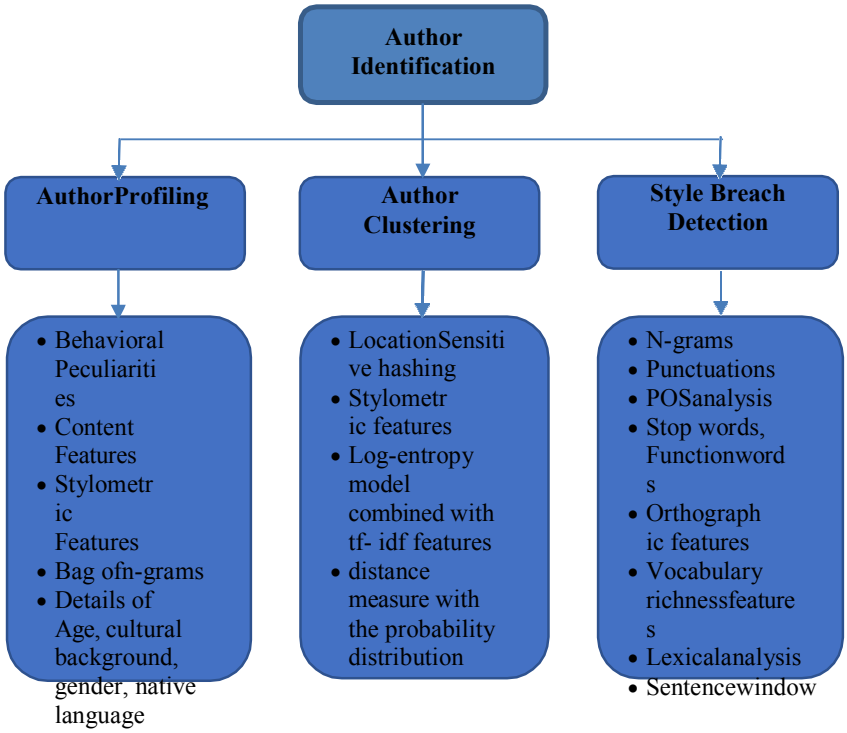
Author Identification
AuthorProfiling
Author Clustering
Style Breach Detection
• Behavioral Peculiariti es
• Content Features
• Stylometr ic Features
• Bag ofn-grams
• Details of Age, cultural background, gender, native language
• LocationSensiti ve hashing
• Stylometr ic features
• Log-entropy model combined with tf- idf features
• distance measure with the probability distribution
• N-grams
• Punctuations
• POSanalysis
• Stop words, Functionword s
• Orthograph ic features
• Vocabulary richnessfeature s
• Lexicalanalysis
• Sentencewindow

**Figure 1:** State-of-art techniques for Author identification and style breach detection

## Relatedwork

As summarized in figure 1, many researchers have come up with studies on identification of authorship among documents with great content similarity [1]. A measurable experiment involving crowd-sourcing, was executed. Authors state that this task is quite challenging. Further, an investigative data analysis on the results of the studies is recommended. Comparison of content features and stylometric features leads to improved algorithms for automatic authorship attribution as well as plagiarism detection. They further contribute in creation of profiles of writers and support intelligence applications to analyzedestructive and bullying messages. This study further could help journals editors also to enforce their journal specific writing styles onauthors.

To recall some earlier studies on Authorship analysis (AA) [2] based on writing characteristics of authors out of segments of huge textual data. These researchers extract an author's individuality and social behaviour peculiarities built on the replicated writing flairs in thetext.

Authorship analysis finds potential application in for research domains, such as cybercrime investigation, psycholinguistics, political socialization, etc. The crucial step of feature engineering process is important part of implementation. Subsequently, the selection of features plays major role in deciding efficiency of the system.

Authors endeavor to simulate the human sentence formation method using a deep learning; and further plan to integrate different groups of linguistic features into disseminated depiction of words in order to learn at the same time the writing style representations based on unlabelled corpus for authorship analysis.

The performance of approach on the problems of authorship characterization and authorship verification with the Twitter, novel, and essay datasets are evaluated by Authors [2]. The experiments suggest that the proposed text representation gives better results than the bag-of-lexical-n-grams, Latent Dirichlet Allocation, Latent Semantic Analysis, PVDM, PVDBOW, and word to vector representations.

This survey aims to prepare a foundation towards spotting locations within a corpus where the authorship deviates, i.e., where the writing flair changes. Consequently, it is diligently associated to all arenas within stylometry, exclusively inherent plagiarism exposure [3] Numerous methods are available that involve lexical feature extraction for creating stylistic fingerprints such as character n-grams [4, 5] word frequencies [6] or word/sentence lengths, syntactic features such as Part-of-Speech (POS) tag structures, structural features such as average paragraph lengths, or indentation usages[7].

As far as multiple-author documents are concerned, overall, interrelated work is infrequent. There are numerous existing methodologies for the text segmentation tasks, in which a text is divided into discrete portions of diverse topics, only few methods target segmentation by other criteria, especially not by authorship. An approach that employs stylometry to automatically detect boundaries of authors of collaboratively written texts is studied [8]. This research does not clearly disclose multiple authors, but delivers clues, such as can be standardized global style of collaboratively written corpus. Additional approaches include neural networks with several stylometric features, a stochastic model on the occurrences of words to split a document by authorship [10].

Lately, a supervised approach, that hybridizes TF.IDF illustration of the documents with features particularly concocted for the task and which makes predictions using some collaborative models including Support Vector Machine, Random Forest etc [11].

Most recently, for English and Urdu text, authorship identification is experimented using the Linear Discriminant analysis model with n-grams texts of authors and cosine similarity [12]. Efficient use of similarity metrics to recognize several learned illustrations of stylometric features to identify the writing style of a particular author is found. The researchers recommend instance-based and profile-based groupings of an author’s text. It is noted that, LDA appropriately handles high-dimensional and sparse data by allowing more communicative demonstration of text corpus. The computational approach levers the heterogeneity of the dataset, multiplicity in writing, and the intrinsic uncertainty of the Urdu language[13].

Another recent approach to author profiling found that identifies various particulars of the author of the corpus such as age and gender [14]. This paper describes a collective task on Bots and Gender Profiling in two languages: English and



DOI: 10.5373/JARDCS/V12SP6/SP20201026
*Corresponding Author: Rajesh Shardanand Prasad E-mail: rsprasad_viit@yahoo.com
Article History: Received: Mar 20, 2020, Accepted: June 22, 2020

Spanish. For example, consider a Twitter feed; the system determines whether its author is a robot or a human. For human authors, it identifies her/his gender. The pre-processing methods, extracted features and machine learning techniques contribute a great accuracy reported as 84.8%.

This survey while discussion with majority of work on English, hereby initiate research performed on regional language of India, i.e. Marathi [15]. For authorship Identification amongst a given set of suspected candidates witnessed a large amount of research that is already been done for the English language. Relatively, fewer researches are carried out for Indian regional languages (e.g. Tamil, Telugu, Bengali and Punjabi) and no evident experimentation is available for Marathi. A set of fine-grained lexical and stylistic features for the analysis of text is used to develop statistical similarity model and SMORDT-Sequential minimal optimization with rule- based Decision Tree approach. It is therefore observed that feature extraction methods used by them show consistent significance in this experiment. Furthermore, it can be noted that the conventional measures such as Recall, Precision, F-measure and Accuracy are still useful to evaluate the performance.

While taking this discussion ahead, it is being noted that machine learning techniques have dominated the field of authorship identification task. Few advancements using deep learning methods is found that define a suitable characterization of texts that depicts the writing style of authors [16]. Deep learning has been lately used in different natural language processing tasks,it has been uniquely used in author identification [16]. Deep learning is used for feature extraction of variable size character n-grams. A Stacked Denoising Auto-Encoder (SDAE) is found useful in extracting document features with different settings, followed by an SVM classifier. The deep learning method is seen to beat its counterparts.

The study so far underlines the significance of feature extraction on the performance of the author identification system. This study therefore gives weightage to Linguistic features for author identification based on the writing style of a particular author. The major challenge involved is the issue such as surface level changes of the author's writing style which should be considered in the domain of authorship verification. This domain is a potential topic for the computer science researchers on authorship verification in the field of computerforensics.

Again, coming back to the central focus of this paper that is style breach detection, a technique [18] that detects style breaches for unknown number of authors within a single document in English is reviewed. Style breaches and mark text boundaries are detected by an unsupervised approach on the basis of different stylistic features. In this technique, features designating common English word frequencies within sentence text, sentence expression and sentence attitude are main focus. These characteristics may not be directly associated to author's style of writing but to the domain of sentence under assessment. An efficient exploitation of sentence window for style detection is found useful. The sentence window may be extended to neighboring sentences during its unsupervised analysis.

While continuing this study for author attribution there is slight a variety of research carried out for distinguishing bots and human beings with gender based on tweets during a Profiling task at again at PAN 2019 [22]. This approach amalgamates character and word n-grams as features for each class trains an SVM classifier for detection task that performs exceedingly well.

## Popularly Used Data Stes

Different datasets are exploited to train and test author attribution models developed so far. One is Reuters 50 50 news dataset that is widely used for authorship identification. Next, Gutenberg story dataset established by Stanford researchers Chen Qian, Tianchang He and Rao Zhang [11]. Further, some author-specific documents representations from PAN12 and Urdu Corpus are found [12].

A training dataset CLEF 2017 [18] containing 187 English text documents of different lengths and sizes over different topics like hotels politics, travel, ,biography etc. had been used for style breach detection,. The uniqueness of this dataset isthat, a truth file is provided, along with each text document which encompasses exact positions of character signifying style violation occurrences within that document, theme of text yet remainsunaffected.

One of the critical point of research contributing a novel task of Native Language Identification (NLI) based on their second language is notified [19]. NLI is a vital application in different extents such as social media investigation, authorship identification, second language extraction, and forensic study etc. An emerging research area of Hyper-dimensional Computing (HDC) is evolving for upcomingresearchers.

For the system authors are planning to develop based on the reviewed papers will organize the task of breach detection and author detection in a multi-authored text into two authorial phases. To do this, two sub- tasks will be accomplished. (1) style breach exposure, i.e., the organizing of a text into stylistically analogous parts, and (2) author clustering, i.e., the organizing of paragraph-span texts by authorship. The similar work is presented and evaluated with elaborated strengths and weaknesses [20].

For the evaluation of character n-gram and lexical features on Spanish text, the corpora are limited; this puts constraints on the amount of research done for this language. Some researchers [21] have built a corpus composed of news articles Argentinian, Mexican, Colombian, Chilean, Venezuelan, Panamanian, Guatemalan, and Peninsular Spanish(which are eight varieties of the Spanish language).This corpus [21] comprises of more than 5000 bulletin documents authored by more than 200 different people (more than 2,000 articles written by male authors and 2,000 by female authors) and includes eight varieties of Spanish.

This investigation on different datasets leads to some uncovered facts. It is observed that few authors have provided opendatasetsto readersalongwiththeirexperimentaldetails[23].Thesolepurposeistopermitfordirectduplication of their findings and reassure assessment andstandardization.

## Discussion Based On Robustreview

A vision to achieve good results with the help of 2,000-word unigrams seems to be achievable. Authors aim to exploit pre-processing methods with the use of around 2,000-word unigrams, and 1,000-word bigrams with utmost hops of 2 words approximately after the survey of state-of-art techniques reviewed herewith. The feasible pre-processing method that the authors plan to undertake will be a mixture of around 1,000-word unigrams, 1,000-word bigrams, and 3,000-character trigrams based on the review.

This review will help to baseline the feature extraction process, by incorporating some traditional stylistic features like Part of Speech Tagging examination and sentence lexical analysis. The sentence window for style detection could be implemented and could be possibly extended to adjacent sentences during its unsupervised analysis.

Furthermore, as far as an important issue of linguistic features is concerned, the features such as stylistic, syntactic, and semantic features could be experimented to improve the accuracy of author authenticity. As a preliminary experiment, authors plan to use Naïve Bayes multinomial classifier to implement the classification prototype for Author Verification.
The major challenges and constraints faced by various researchers whose literature is reviewed herewith, includes issues regarding locating boundaries, where author's style changes. Furthermore, for multiple authors, difficulties are reported regarding, establishing an optimal number of clusters, cluster homogeneity, cluster completeness etc. This paper aims to find solution to overcome these challenges.

The discussion so far indicates that the task of analyzing Natural Language that in turn leads to the solution to Language identification problem has been widely investigated for over five decades. Today, Language identification has become a crucial measure of numerous text processing tasks, this study further explores features and approaches implemented for Language Identification. At the outset, authors ascertain exposed issues; review the work to date on each issue, and chalk-



DOI: 10.5373/JARDCS/V12SP6/SP20201026
*Corresponding Author: Rajesh Shardanand Prasad E-mail: rsprasad_viit@yahoo.com
Article History: Received: Mar 20, 2020, Accepted: June 22, 2020

out future directions for research in style breach detection. Figure 2 depicts typical Natural Language processing steps in order to solve the underlying problem of Language identification. The process of style breach detection starts with processing of words listing that leads to sentence segmentation and analysis such as Lexical, Content specific, behavioral and style labeling.

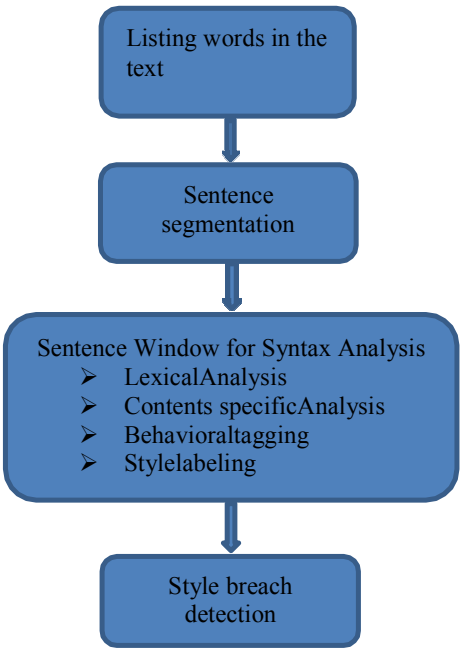


**Figure 2:** Planned outline for Author Breach Detection

## Futuredirection

Our plan as future work is to consider the author's institution into account to serve as anaspect in the paper selection process—similarly with papers published in the same journal.We also look to investigatethe text passages written within span of years by the same author. We further wish to extend our work by including perspective of human annotators, properties of the thinking process for growing and diversifying the set of experiments.

Future research will focus on exploring better models to capture writing styles and proposing models for other languages. Currently the representation learning models are simple one-layer neural networks. A recurrent neural network (RNN) with long- short term memory is more suitable for capturing the contextual relationship over long text.

For learning the syntactic modality representation, a recursive neural network that operates on the fully parsed syntactic tree will fit more into the nature of grammatical variations than the current one. Moreover, this work only focuses on capturing the variations in English writing. Additional changes need to be applied for text in other languages.

This research could also be extended in one of the evolving era of HDC, where high dimensional vectors representing feature vectors that are also known as hyper vectors could be employed. The basic idea is to break the character streams in bi-grams and tri-grams which are used for generating comment hyper vectors. These hyper vectors will be then joined to generate diff erent language profile vectors. Profile hyper vectors will then be used for classification of test comments. These techniques are anticipated to validate around 60 -70% of training instances with the accuracy in the range of 70 -75% based on the reviewedliterature.

## Conclusion

The above sections have illustrated the experiences of various researchers to sufficient depth. It is also to be concluded based on this study that some words need to be treated specially to improve the accuracy. These words are extensively used nowadays on social media conversations and depict peculiarity of authors, such as tc (take care), gn (Good night), gm (Good morning), hbd(Happy Birthday), ty (thank you), pls(please), u (you), y (why), r (are), sry (sorry), fyi (for your information), etc. Another important reading is about punctuation marks that play a significant role in natural language analysis. In addition, with the development of modern methods for style breach detection, the conventional evaluation measures such as training effort, precision, recall, accuracy are still considered to be an irreplaceable criterion for performance and efficiency calculations. Furthermore, the system performance could be improved by introducing more



DOI: 10.5373/JARDCS/V12SP6/SP20201026
*Corresponding Author: Rajesh Shardanand Prasad E-mail: rsprasad_viit@yahoo.com
Article History: Received: Mar 20, 2020, Accepted: June 22, 2020

stylometricindicators.

## Acknowledgment

Author acknowledges the support rendered by the Management and colleagues of parent institute –Sinhgad Institute of Technology and Science (under SavitribaiPhule Pune Univeristy), Pune, India‖. He also acknowledges the support by all faculties and management of Lincoln University College, Malaysia.

## References

[1] AndiRexha, Mark Kröll, Hermann Ziak, and Roman Kern, Authorship idevtification of documents with high content similarity, In Springer, Scientometrics, Scientometrics. 2018; 115(1): 223–237, doi: 10.1007/s11192-018-2661-6,2018.

[2] Ding SHH, Fung BCM, Iqbal F, Learning Stylometric Representations for Authorship Analysis, Cheung WK, IEEE Trans Cybern. 2019 Jan;49(1):107-121. doi: 10.1109/TCYB.2017.2766189. Epub 2017 Nov21.

[3] Zlatkova, D., Kopev, D., Mitov, K., Atanasov, A., Hardalov, M., Koychev, I., Nakov, P.: An Ensemble-Rich Multi-Aspect Approach Towards Robust Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org(2018).

[4] Noreen, E.W.: Computer-Intensive Methods for Testing Hypotheses: An Introduction. John Wiley and Sons,1989.

[5] Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609.

[6] Kestemont, M., Luyckx, K., Daelemans, W., Crombez, T.: Cross-genre authorship verification using unmasking. English Studies 93(3), 340–356 (2012).

[7] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276, 2007.

[8] Mike Kestemont, Michael Tschuggnall, EfstathiosStamatatos, Walter Daelemans,Günther Specht,2 Benno Stein,4 and Martin Potthast5Cross- domain Authorship Attribution and Style Change Detection, Overview of the Author Identification Task atPAN-2018.

[9] Dimitrina Zlatkova_1, Daniel Kopev_1, Kristiyan Mitov_1, Atanas Atanasov_1, Momchil Hardalov1, Ivan Koychev1, and PreslavNakov, An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection Notebook for PAN at CLEF-2018, Vol.2125.

[10] Giannella, C.: An improved algorithm for unsupervised decomposition of a multi-author document. Technical Papers, The MITRE Corporation, February2014.

[11] Daniel Kara´s, Martyna ´Spiewak and PiotrSobecki, Author Clustering and Style Breach Detection, OPI-JSAatCLEF2017, Notebook for PAN at CLEF2017.

[12] Waheed Anwar , Imran SarwarBajwa, and ShabanaRamzan, Design and Implementation of a Machine Learning-Based Authorship Identification Model, Scientific Programming Volume 2019, Article ID 9431073, 14pages.

[13] Yaakov HaCohen-Kerner, Natan Manor, Bots and Gender Profiling of Tweets using Word and Character N-Grams Notebook for PAN at CLEF 2019 1, Conference: The Tenth International Conference of the CLEF Association (CLEF 2019)., At Lugano,Switzerland.

[14] Yaakov HaCohen-Kerner, Daniel Miller, YairYigal, ElyashivShayovitz, Cross-domain Authorship Attribution: Author Identification using Char Sequences, Word Unigrams, and POS-tags Features Notebook for PAN at International Conference on Cross-domain Authorship Attribution, at CLEF2018.

[15] Kale Sunil Digamberrao, Rajesh S.Prasad Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi, Elsevier Procedia Computer Science. Volume 132, 2018, Pages1086-1101.

[16] Ahmed M. Mohsen; Nagwa M. El-Makky ;NagiaGhanem, Author Identification Using Deep Learning, 15th IEEE International Conference on Machine Learning and Applications (ICMLA), , INSPEC Accession Number: 16651340, DOI: 10.1109/ICMLA.2016.0161,2016.

[17] Bhanu Prasad A., Rajeswari S., Venkannababu A., Raghunadha Reddy T. (2018) Author Verification Using Rich Set of Linguistic Features. In: Satapathy S., Tavares J., Bhateja V., Mohanty J. (eds) Information and Decision Sciences. Advances in Intelligent Systems and Computing, vol 701. Springer,Singapore.

[18] Jamal Ahmad Khan, Style Breach Detection: An Unsupervised Detection Model Notebook for PAN at CLEF2017.

[19] AshishPatelandPratikShah,Hyper-dimensionalComputingfor IndianNativeLanguageIdenti_cation,CEUR,Volume1-2266,IIITV@INLI,2018.

[20] Walter Daelemans, GüntherSpecht, Benno Stein, and Martin Potthast, Style Breach Detection and Author Clustering, International Conference of the Cross-Language Evaluation Forum for European Languages , CLEF 2017: Experimental IR Meets Multilinguality, Multimodality, and Interaction pp145-151.

[21] Sanchez-Perez M.A., Markov I., Gómez-Adorno H., Sidorov G. (2017) Comparison of Character *n*-grams and Lexical

Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. In: Jones G. et al. (eds) Experimental IR Meets Multilinguality,Multimodality,andInteraction.CLEF2017.LectureNotesinComputerScience,vol10456.Springer.

[22] MahendrakarSrinivasarao and Siddharth Manu, Bots and Gender Profiling using Character andWord N-Grams, at International Conference CLEF 2019, 9-12 September 2019, Lugano, Switzerland.



DOI: 10.5373/JARDCS/V12SP6/SP20201026
*Corresponding Author: Rajesh Shardanand Prasad E-mail: rsprasad_viit@yahoo.com
Article History: Received: Mar 20, 2020, Accepted: June 22, 2020