Journal of Cases on Information Technology
Volume 24 • Issue 4

State of the Art in Authorship Attribution With Impact Analysis of Stylometric Features on Style Breach Prediction

Rajesh Shardanand Prasad, MIT Art, Design, and Technology University, Pune, India*
Midhun Chakkaravarthy, Lincoln University College, Malaysia
https://orcid.org/0000-0002-0107-885X

ABSTRACT

The most influential research was studied that spans the domains from authorship attribution and stylometry. The reference material contributes robust classifiers with reasonable array of feature extraction techniques, such as Dirichlet-multinomial change point regression to extract the progress of inscription elegance with time, comprising plodding variations in stylishness as the author ages and unexpected vicissitudes. This paper presents quantifiable evaluation of the research in terms of year-wise research output, diversity of applications, nature of collaboration, characteristics of highly productive techniques, and the benchmark of performance criteria by eminent high impact researchers. The outcomes of this study can by deployed for dialectology analysis and corpus linguistics, stylistics, natural language processing, classification, literary and historical analysis, and forensic analysis.

KEYWORDS

Authorship, Linguistic Feature, Stylometric

INTRODUCTION

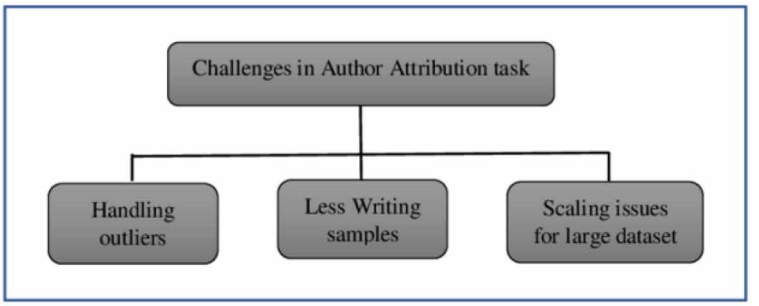
Authorship Attribution aids to recognize the right author of a specified unnamed article from a set of contender authors. Authors have presented a rigorous exploration of state-of-art methods in this regard already (Rajesh Shardanand Prasad, Midhun Chakkaravarthy, 2020). The applicability of this task can be originated in numerous areas, for example law enforcement interventions and data storage and retrieval. These application domains are not restricted to a explicit language, communal, or culture. However, most of the prevailing solutions are intended for English, and a slight consideration has been rewarded to regional languages.

Figure 1 shows the major challenges to apply this author attribution research to efficiently and effectively to regional languages or languages other than English. They are listed as follows:

- Handle outliers in the dataset;
- Scale when the size of the candidate authors set increases; and
- Perform well when the number of writing samples for each candidate author is low.

DOI: 10.4018/JCIT.296716
*Corresponding Author

Figure 1. Major challenges before author attribution researchers



The first and foremost step includes identification of stylometric feature space. Further tasks of feature extraction and identification may employ state-of-art classifiers.

Most of the research literature review and their evaluation emphasize on the effectiveness of each stylometric features. The major findings of this study state that combination of all groups of the stylometric features outclasses the contemporary combinations. To conclude, authors cross equates the feature spaces and classification approaches of all systems. This literature evaluation concludes that Hybridization of stylometric features enhances the performance of the system increases with the increase in the number of candidate authors. Furthermore, such hybridized feature space provides improved efficiency than the feature space used by the non-hybridized feature space.

RELATED WORKS

Authors hereby begin the review with state-of-the-art and vigorous method to stylometric investigation starved of explanation and grasping etymological and sub-etymological evidence (Hou, Renkui & Huang, Chu-Ren., 2019). Especially, authors suggest to influence the lingual information of qualities and rimes in Mandarin Chinese mechanically pulled out from annotated texts. The texts from dissimilar writers were characterized by manners, manner themes, and expression span themes as well as rimes and rime themes. This methodology ensures effective results in author stylometric analysis with the use of Support vector machines and random forests.

Exploring Some Untraveled Paths

Another research that deliberates some uncommon approaches of author attribution scheme in Bengali fiction deploying stylometric approach (Hossain, Anika, et.al. 2020). System scrutinizes whether it is conceivable to recognize the genuine writers of some unidentified Bangla papers by means of a Machine Learning (ML) algorithms and Artificial Neural Networks (ANNs). The classifier votes contribute a decision for a dataset archived from various political writers. This research shows that the Multilayer feedforward neural network and SVM classification model are prove to be an established option to develop an attribution system paired with a voting system.

Authors scholarly research (Zhu, et. Al. 2020), presents a reckonable examination of style and section in the classical Chinese novel, Dream of the Red Chamber (DRC), and debate the insinuations for the unclear composition of the novel. Initially, authors scrutinize the quantity of section in across the sections of DRC. To do this they deploy Principal Component Analysis (PCA). This stylistic analysis of the prose portions of the novel produce novel and stimulating significances, which proves that stylometric apparatuses can be used to enable complex researches of traditional Chinese literature.

Furthermore, Function word content investigation (Research note 2020) is a confirmed technique beforehand to explore doubtful composition in papers for example past literatures and narratives.

Similar techniques are also demonstrated to evaluate Supreme Court Justices' lawful emotions, representing over-all involvement of clerks and dissimilarity associated to their impact.

Recent international and National happenings (Belvisi et. Al. 2020) have demonstrated that, communications and writings forwarded on the Internet are utilized for felonious inquiries. Regrettably, the authorship of many of them remains unidentified. This is used for identifying authors of tweets using prevalent features experimented conventionally. A corpus of 40 users, with 120 to 200 tweets per user is used and resulted into 92% to 98.5% accuracy.

To locate extensions of applications of Stylometric methods includes revealing resemblances amongst texts and, combined with network analysis, to portray the stylistic relations between those texts for Jacob and Wilhelm Grimm brothers (Gabriela. et. al. 2020).

Next, authorship identification for 547 Thai documents from 200 authors for solution using probabilistic k nearest neighbour classifier is studied (Sarwar. et. al. 2020). Individual documents are transformed into an assortment of point sets. Precisely, this document alteration enables use of distance measures, stylistic variations and multiple estimates for a query document.

Till now authors found effective ML techniques focus on retrieving an individual's distinguishing means of speaking or writing. This exploration showed that each individual author owns a dialect with a significantly large corpus, a tool named as 'JStylo' equipped with SMO classifier is found to be effective (Al-Khatib. et. al. 2020).

Statistical and Linguistic Models

Now authors take this discussion of authorship attribution, forward towards employment of statistical and linguistic methods. Nevertheless, Stylometric authorship attribution, is yet spotted as widely used because of its meticulousness and effectiveness (Omar 2020). Even so, numerous researchers experience unreciprocated difficulties for some languages specifically Arabic. An endeavour based on linguistic peculiarities that are not usually considered in standard authorship systems.

A more important finding here in case of Arabic is that, the morphological features represent exclusive stylistic features that can be usefully used in evaluating authorship in debated manuscripts and literatures. The supposition is that much of these morphological features are misplaced due to the implementation of curtailing. Authors will plan the implementation by concerning the ill-effects of stemming on system accuracy and performance.

Next, a significant point of terminology lushness articulated by–token ratio, a parameter of text motion, mean word length, mean verb distance, and cluster analysis of the most frequent words cannot be overlooked (Miroslav 2019). Evaluation of these parameters help to examine author-specific differences and assessment of individual styles.

While looking at stylometric evaluation, there is an attention seeking issue rising up through Natural Language Research, that deals with artificial binge of fake news or information (Schuster 2020). This issue is of serious concern for upcoming researchers. A solution to this problem can be achieved by recognizing text provenance, it fails to distinguish legitimate LM applications from those that introduce false information. Two standard models to verify resemblance between nasty and genuine sample texts, used in auto-completion and editing-assistance sceneries. This climax the prerequisite for non-stylometry methods in perceiving malware engendered fake news fabrication.

Another interesting tweet-based author identification using n-grams and word n-grams that treats the single-labelled multi-class problem in order to identify author. This study (Tarmizi 2020) demonstrates outstanding deployment of SVM over Naïve Bayes classifier.

Apart from author identification, few studies raised questions about, authentication and analytical information of authors. A recent approach that generates analytical profile of an author using attributes such as age, gender, and location by investigating their panache of inscription. These further employs combination of syntactic features and content-based features (Reddy 2020). Liu (2018) lengths, dispersion of word lengths, dispersion of sentence lengths, part of speech (POS), POS of content words, POS of function words, punctuation, high-frequency words, n-gram of POS, n-gram of words,

n-gram of punctuation and multiple features. Unique implementation of hierarchical clustering demonstrates profound variation over the years for the same author.

Authorship Confirmation Tasks

Authorship confirmation is the job of extracting if two texts were authored by the single writer on the basis of a writing elegance examination. A study talks about author mystification as the confrontational job of averting an effective corroboration by changing a manuscript's bravura so that it does not appear to be similar to that of its genuine writer any longer (Bevendroff 2020). Authors observes another algorithm that does the same tasks on a wide-ranging assessment to establish the virtues of the most recent techniques of writer confirmation to endure obfuscation. The two noteworthy techniques found replicating inscription style variance as the Jensen-Shannon distance amid the character n-gram distributions of manuscript that further employs an author's scripting style in a urbane way by means of heuristic search. This approach, highlighting mystification task, researchers deploys the enormous domain of word-based variations so as to discover a summarized form of the to-be-obscured text that has a satisfactorily huge Jensen-Shannon detachment at negligeable outlays in standings of text semantic forfeiture. These researchers systematically examined, enumerated, and exemplified the foundation of this method, describe rewording operators, originate manuscript span-invariant thresholds for expiry, and grow an operative obfuscation outline. To note final observation, an experimentally efficient evaluation to a converse obfuscation occurrence in contradiction of our obfuscation approach on top of likely techniques.

This literature review has presented an in-depth review of authorship stylometric attribution techniques. This discussion can be further supported by a unique statistical technique that responds to fictional queries, for example authorship attribution, in a measurable means (Ross, 2019). An ensemble of analytical frameworks grounded on Dirichlet–multinomial alteration point regression which can seizure the progression of scripting elegance over period, encompassing together plodding vicissitudes in panache as the author develops, and sudden variations which can be produced by thrilling proceedings in the writer's lifespan. The motivation behind this research, is found to be based on disastrous Alzheimer's disease of Agatha Christie, and the signals of slow gist in her writings. The implications for stylometry and authorship attribution are discussed.

Author Profiling

Furthermore, this task of author outlining has now been portrayed as a data-centric computational dialectology which is casted to forecast the demographic contours such as age, gender, location, local language of authors by dispensation of the word-based contents of their inscribed manuscript (Kavuri, 2020). Author profiling is thus prevalent in current existences owing to its latent submissions such as advertising investigation lessons, criminological linguistics, sanctuary and mythical research. Next, a widespread exploratory area called Stylometry is the major focus of this study [18]. The foremost attentiveness of this effort is on feature engineering, the development, evaluation, and application of the feature set in the context of machine learning techniques to author profiling. These stylistic topographies attained the finest exactitudes for forecasting gender when likened with state-of-the-art methods in numerous situations, particularly when amalgamated with other structures.

Taking this discussion of statistical examination further, let the authors present two more approaches, on the basis of the study of digits manifestation in an intelligible legendary text. The primary method is associated to the examination of the frequency distribution of noteworthy digits of numerals happening in the corpus (Kavuri, 2020). This method is suitable for rapid examination of mutual authorship: the next approach is uncertain if the frequency distributions are adequately dissimilar. Another method is the leeway of the first method and necessitates the analysis of the frequency distribution of the numerals.

So far, this literature review reported representative statistical and linguistic techniques. Although the state-of-the art techniques revolve round accuracy and effectiveness criteria, there has been several

authorship glitches persistent as unsettled in Arabic (Omar 2020). The reasons are observed to be ascribed to diverse influences together with philological individualities that are not typically measured in conventional authorship systems. The exclusive stylistic topographies in Arebic are exploited in assessment of authorship in controversial texts and writings. Mainly this research endeavour investigates the efficacy of stemming in the stylometric applications to authorship attribution in Arabic. The results show that stemming has adverse effects on the correctness of the clustering performance and thus on the dependability of stylometric authorship assessment in Arabic. The unusual stylistic topographies of the attachment procedures in Arabic can, consequently, be practically utilized for refining the performance of authorship attribution applications in Arabic.

An altogether new application area for a related application of author contents attribution is found (Iyer 2020). Selecting an apt conference and journal for a research is vivacious as it aids in accomplishment of the correct addressees and likewise to supplementary fortuitous of getting their research work printed. The issue of endorsing suitable conferences to the writers to upsurge their probabilities of acceptance. The noteworthy contribution involves application of social network of the authors and the content of the paper in the settings of dimensionality reduction and topic modelling. Correspondence Analysis (CA) springs suitable associations between the entities in question, such as conferences and papers. This method overperforms the conventional content-based filtering, collaborative filtering and hybrid filtering.

There are many situations when a well-known author writes with some unknown author. The task of authorship identification is practically approached utilizing the minimum arithmetic mean of errors on the basis of joint selection of the strictures is found (Iskhkova 2018) for Russian and Arabic.

Stop Words and Stemming Issues

There is an interesting conclusion made by few studies mentioning that instead by neglecting stop-words, a modest worldwide thresholding plan to create virtual links Word co-occurrence networks have been active to examined texts both in the real-world and hypothetical situations (Quispe 2013). This can be regarded as the finest application of word embeddings as a tool to create virtual links in co-occurrence networks whitethorn enhance the excellence of classification systems.

An innovative application of a Zipf's law that describes an inverse proportion amongst a term's grade and frequency in a specified manuscript and, coarsely distributing the vocabulary to recurrent words and occasional ones (Mokryn 2020). This technique, named as Latent Personal Analysis (LPA), seems great for discovery of such domain-based personal signatures.

The same team of authors, further extend this research with a Latent Personal Analysis (LPA) (Mokryn 2020), for discovering domain-based personal signatures. This research work identifies words contributing to the dissimilarity amid a handler's terminology from the domains. It is commendable that this LPA method works significantly well for two kinds of caricature in social media: (1) authors with numerous accounts; (2) forward-facing-user accounts, operated by several authors thus useful in capturing individual autographs in an extensive variety of methodical provinces in which the voters have a wide-range spreading of rudiments.

The researchers in authorship stylometry have witnessed the significant role of Semantic web or WordNet. On the similar lines, the Word Adjecency Network (WAN) for authorship detection or attribution for a predominent language such as English to express the associations between words via additional helper words called as function wordsincluding prepositions, articles, conjuctions, pronouns, auxilary words, is shown (Segarra 2020). This study shows that the dependence among these function words may undermine the authorship attribution method.

Furthermore, Binary authorship attribution that denotes to the development of evidence linked to the writer(s) of nameless binary code on the foundation of stylometric features mined from the code (Alrabaee 2020). Though, in reality, authorship attribution for binary code yet necessitates substantial physical and fault-disposed to reverse engineering examination, which can be an intimidating chore assumed the utter capacity and involvedness of present group of malicious software. This paper

presents a unique compiler-unconvinced technique for recognizing the writers of program binaries. The peculiarity of this method imprisonments an author's programming and coding traditions.

In addition to regular and conventional endeavours to author style capturing, a really interesting uncommon area of research is discovered during this survey. Phylogenetics that deals with classification of a group of languages with reference to their ancient progression (Gamallo, 2020) identified some out-of-the-way languages and quantified the detachment amid them and from the rest of the European languages, in order to throw light on the philological detachments and immediacies of these provocative languages without considering phylogenetic issues, during the experiments performed on 40 European languages.

Sentiment uncovering and cataloguing is the newest craze for social analytics on Net. With the collection of real-world applications in medicine, economics, mass media, shopper souks, and administration, extracting the power of speech of community to get intuition to mark evidence and appraisals is non-trivial. With a noticeable upsurge in the extent, prejudice, and assortment of social internet-data, the indistinctness, ambiguity and fuzziness inside the statistics has amplified assuredly. The Soft computing practices (Kumar 2019) are found to face this fuzziness in practical applications such as twitter.

This paper after considering the WAN, now explores the latent capability of another of Bootstrap Consensus Networks (BCN)-yet unique style to produce conceptions in stylometry by charting resemblances of writer style between manuscripts into the arrangement of a network-for extensive authorship attribution responsibilities (Meier 2020). This extensive experimentation, suggests that the scope of the erected networks be contingent deeply on the kind of variables and distance measures used.

After reviewing the handling of Arabic issues, there is yet another extension of research finding laid (Al-Sarem 2020) utilising, the Technique for Order Preferences by Similarity to Ideal Solution (TOPSIS) technique to pick the primary classifier of the collaborative approaches. To achieve this, they use attribution geographies, bunch of stylometric topographies and diverse terms via numerous apparatuses have been mined. The algorithms employed include Adaboost and Bagging ensemble methods involving Arabic Fatwa datasets that claimed to enhance the performance to great extent.

With the enormous increase in Internet data, there is an increasing need to address the cyber-crime also. A research that addresses term importance by using term weight measure, used to calculate the document weight is evident (Reddy 2020). Furthermore, the issue of Plagiarism, which is a foremost task of speculative deceitfulness; henceforth the discovery of text-stealing is very vital. Consequently, Copy Recognition is a flourishing domain of investigation in Natural Language Processing that includes the credentials of embezzled sections of manuscript and the recovery of the foundation of the inventive text. Plagiarism detection based on Word2vec, Monte Carlo ANN, Candidate Retrieval and Text Alignment, PV-DM and PV-DBOW, Rabin-Karp Algorithm, IR-based plagiarism detection, LSI, and Joint Word Embedding is studied (Narayanan 2020). Refer Figure 1 for details.

It is evident that sophisticated Deep Learning methods enhance accuracy than other techniques. Finally, it is stated that there is an immense need of Cross-Language Plagiarism Detection.

Now the discussion has arrived at the clustering techniques employed in the latest period, that necessitates mentions Vector Space Clustering (VSC) procedures in the computational scrutiny of mythical data together with type classification, leitmotif investigation, stylometry, and authorship attribution. Despite the success of VSC methods these techniques suffer from the delinquent of feature selection task, that requires extraction of the most idiosyncratic topographies within a corpora that is handled with term weighting approaches along with TF-IDF, and Principal Component Analysis (PCA) (Omar 2017).

An attention catching research involving Principal Component Analysis (PCA) and Factor Analysis (FA) is functional to group thirteen dissimilar prearranged groups about the Suras of the Holy Quran (Wang 2012). Furthermore, A research resource is emerged as an ensemble of reliance syntax trees of demonstrative writings from ancient Greek prose authors is used

for pedagogical purposes (Gorman 2020). This further enables research in dialectology investigation and corpus linguistics, stylistics, natural language processing, classification, and literary and historical analysis.

Social Media and Author Attribution Issues

Social media is found as a promising source of analysis available to the eminent researchers. A new-fangled multi-modal writing substantiation tactic for social media texts is found (Suman 2020) in Facebook and twitter through emojis and other expression components. This analysis is based on the writing style of a user word choices, sentence structures, usage of punctuation symbols, and use of emoji can make a huge difference. A multi-modal Siamese-based framework for involuntary mining of structures from the specified typescripts and emojis.

There is another application of historical studies supported by author style analysis that extends application of social media data (Juola 2020) through representative case study analysis. An interesting fact that the author sticks to his/her Native Language while developing manuscripts. Native language of an author is an inseparable part of his/her writings. This area seems to be really exciting and is termed as Native Language Identification (NLI). This is accomplished by analysis of the content-specific/social-network landscapes, language-convention-outlines (Sarwar 2020). This technique uses the probabilistic k nearest neighbours' classifier.

Table 1 lists some promising techniques in the domain of Author Attribution and style breach prediction.

Table 1. Promising techniques based of related work

Sr. No.	Technique	Reference
1.	Multilayer feedforward Neural network	4
2.	SVM classification	4
3.	Principal Component Analysis (PCA)	5
4.	Function word content analysis	5
5.	SMO and MLP algorithms via 'JStylo'	9
6.	Statistical analysis	10
7.	Clustering Structures	16
8.	Dirichlet–multinomial change point regression	17
9.	Dimensionality Reduction and Topic Modelling	21
10.	Word co-occurrence networks	23
11.	Latent Personal Analysis (LPA)	25
12.	BinAuthor, the first compiler-agnostic method	28
13.	Bootstrap Consensus Networks (BCN)	29
14.	Technique for Order Preferences by Similarity to Ideal Solution (TOPSIS) Vector Space Clustering (VSC)	31
15.	Glove, Word2Vec and FastText	33
16.	Word Adjacency Networks (WAN)	35
17.	Probabilistic k Nearest Neighbours Classifier	39

RESULTS AND DISCUSSION

Authorship recognition helps to categorize the correct author of a specified unidentified article from a group of contender authors. The usefulness of this job is evident in numerous areas, for example law enforcement agencies and information retrieval. These application areas are not restricted to a definite linguistic, masses, or civilization. Nevertheless, most of the prevailing solutions are premeditated for English, and a slight consideration has been rewarded to languages like Chinese, Korean and Thai. These current implementations are not straight pertinent to these languages owing to the dialectal alterations amid these two languages.

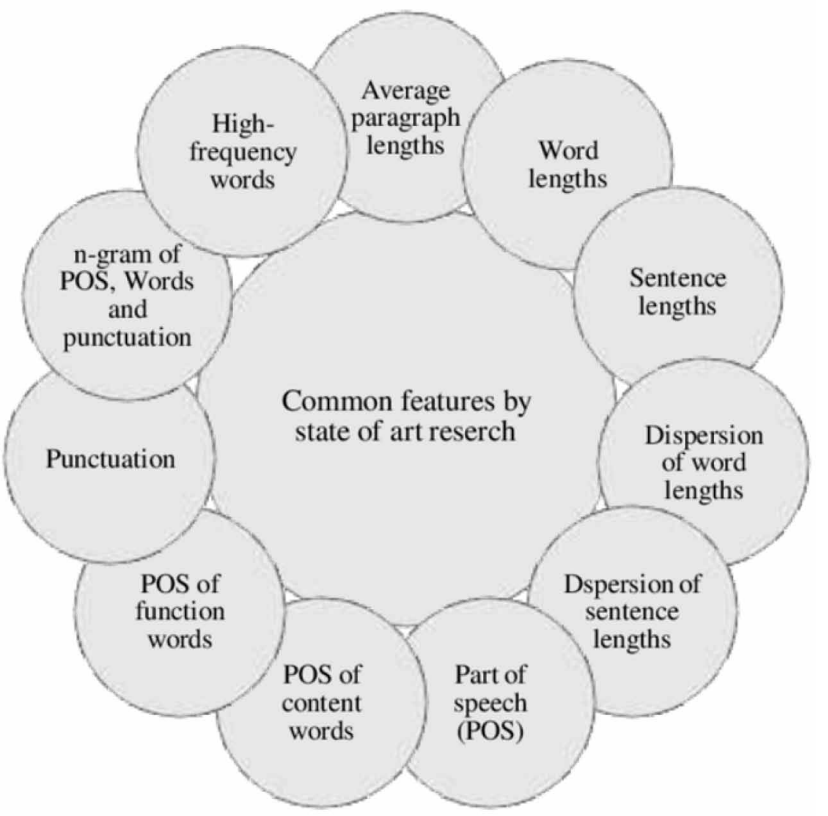
Furthermore, the present implementation designed for the above-mentioned languages is capable of:

1. Leveraging outliers in the dataset;
2. Gauge up in case of increase in the size of the candidate authors set; and
3. Improved performance even with low number of writing samples for each candidate.

Figure 2 depicts the commonly identified state-of-art techniques and set of stylometric features used in them and their impact on the efficiency. Figure 2, depicts contribution of suggested style feature combination on efficiency with three outcomes.

Authors depict a stylometric feature space for the authorship attribution job. Grounded on the stylometric feature space, authors hereby propose an authorship recognition and attribution implementation that exploits the set of few algorithms listed as probabilistic k nearest neighbours'

Figure 2. Some common Style features that can be used in combinations



classifier, Support Vector Machines, Clustering algorithms, Multilayer Perceptron and Random Forests. Through transformation of respective essay into an assortment of opinion sets.

Precisely, this text conversion allows us to:

1. Exploit group of distance measures related with an outlier management apparatus;
2. Depict style disparities inside a document; and
3. Yield manifold guesses for an interrogation or test document.

Authors generate a novel authorship identification framework containing 600+ documents from 100+ authors, which is meaningfully greater than the quantity of document dataset through the current work. The experimental results demonstrate that this analysis provides insights into the constraints of the existing implementations and dominates all contemporary solutions that exceeds accuracy levels beyond of 90%. Furthermore, authors examine the efficiency of individual stylometric features group with the assistance of an extirpation learning. Author explored that combination of all classes of the stylometric features outpaces the other combinations. To understand the effect of different features, authors ranked the selected features based on their information Gain Ratio (IGR). IGR of a feature F_i is defined as:

$$IGR(F_i) = (H(D) - H(D|F_i))/H(F_i)$$

where D is document class and H is entropy.

Lastly, Authors cross validate the feature spaces and classification methods of all implementations. Authors observe the following facts:

1. This exploration is capable to upgrade as the number of contender authors upsurge;
2. The recommended class of features outperform all the contemporary techniques; and
3. The stylistic feature space delivers improved presentation than the feature space used by the existing study. Refer to Table 1 for experimental results.

CONCLUSION

Authorship Attribution and style breach detection is the upcoming fascinated area for community analytics on Internet. With the range of real-world applications in Medical, economics, mass media, buyer marketplaces, and administration, refining the feedback of public to visualize to goal information and appraisals is negligible. Through a noticeable upsurge in the volume, context, and variety of social media-data, the ambiguity, vagueness and fuzziness inside the literature has amplified at large

Table 2. Comparative analysis of contemporary techniques on efficiency with author and corpus statistics

Techniques	Total Number of Authors	Total Number of Documents	Consolidated Length of Document Word Length	Average Document per Author	Efficiency
Support Vector Machines	5	250	100	65	91.3%
Random Forests	4	150	300	55	95.8%
Multilayer Perceptron	25	950	150	45	93.9%
Clustering Methods	10	300	1000	30	93.3%
Probabilistic KNN	10	3000	120	1600	92.6%

level. Evolutionary Algorithms and techniques have been employed to hold this fuzziness in everyday situations. This research is an endeavour to comprehend the viability, scope and significance of this coalition of utilising Intelligent techniques ranging from author posts analysis on social media accounts, style breach detection to Artificial Text Emulation applications. This paper demonstrates a methodical literature review to assemble, reconnoitre, appreciate and analyse the contributions and tendencies in a well-organized method to recognize investigate breaches defining the future predictions of this connection. The influence of this research is expected to be is substantial as the key emphasis is to analyse and assess the usage of Intelligent practices for Author Stylometry Attribution. Furthermore, as associated to the preceding evaluations this in-depth methodical tactic to categorize, collect experiential indication, construe results, judgmentally analyse, and assimilate the answers of all pertinent superior revisions to handle specific research issue relating to the distinct research areas.

REFERENCES

Al-Khatib, M., & Al-qaoud, J. (2020). Authorship verification of opinion articles in online newspapers using the idiolect of author: A comparative study. *Information Communication and Society*, 1–19. doi:10.1080/1369118X.2020.1716039

Al-Sarem. (2020). Ensemble Methods for Instance-based Arabic Language Authorship Attribution. IEEE Access. doi:10.1109/ACCESS.2020.2964952

Alrabaee, S. (2020). Authorship Attribution. In *Binary Code Fingerprinting for Cybersecurity. Advances in Information Security* (Vol. 78). Springer.

Belvisi, N., Muhammad, N., & Alonso-Fernandez, F. (2020). *Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features*. Academic Press.

Bevendorff, J., Wenzel, T., Potthast, M., Hagen, M., & Stein, B. (2020). On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification. *Information Technology, 62*. 10.1515/itit-2019-0046

Gamallo, P., Campos, J., & Alegria, I. (2020). Measuring Language Distance of Isolated European Languages. *Information.*, *11*, 181. doi:10.3390/info11040181

Gorman, V. B. (2020). Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data*, *6*(1), 1.

Hossain, A., Akter, N., & Islam, M. (2020). *A Stylometric Approach for Author Attribution System Using Neural Network and Machine Learning Classifiers*. .10.1145/3377049.3377079

Hou, R., & Huang, C.-R. (2019). Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, 1–23. doi:10.1017/S135132491900010X

Iskhakova, A., Kruglova, S., Melnikov, S., & Sidorov, E.(2018). *The Approach to Minimize the Impostor Method Errors in the Author Identification Open Problem*. Academic Press.

Iyer, R. R., Sharma, M., & Saradhi, V. (2020). *A correspondence analysis framework for author-conference recommendations.* arXiv preprint arXiv:2001.02669.

Juola, P. (2020, January 1). Authorship Studies and the Dark Side of Social Media Analytics. *Journal of Universal Computer Science*, *26*(1), 156–170.

Kavuri, K., & Kavitha, M. (2020). *A Stylistic Features Based Approach for Author Profiling*. .10.1007/978-981-15-0426-6_20

Kubát, M., Mačutek, J., & Čech, R. (2019). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*. 10.1093/llc/fqz089

Kumar, A., & Jaiswal, A. (2019). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation*, *32*, e5107. doi:10.1002/cpe.5107

Liu, Y., & Xiao, T. (2018). A Stylistic Analysis for Gu Long's Kung Fu Novels. *Journal of Quantitative Linguistics*, *27*, 1–30. doi:10.1080/09296174.2018.1504411

Meier, F., Larsen, B., & Stjernfelt, F. (2020). *Exploring the Potential of Bootstrap Consensus Networks for Large-scale Authorship Attribution in Luxdorph's Freedom of the Press Writings*. Academic Press.

Mokryn, O., & Ben-Shoshan, H. (2020). *Domain-based Latent Personal Analysis and its use for impersonation detection in social media.* arXiv preprint arXiv:2004.02346.

Narayanan, A. (2020). A Survey on Plagiarism Detection Techniques. *International Journal of Psychosocial Rehabilitation*, *24*, 1564–1572. doi:10.37200/IJPR/V24I1/PR200254

Omar, A. (2017). *Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods*. Academic Press.

Omar, A., & Ibrahim, W. (2020). The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic. *International Journal of Advanced Computer Science and Applications.*, *11*. Advance online publication. doi:10.14569/IJACSA.2020.0110114

Omar, A., & Ibrahim, W. (2020). The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic. *International Journal of Advanced Computer Science and Applications.*, *11*. Advance online publication. doi:10.14569/IJACSA.2020.0110114

Prasad, R. S., & Chakkaravarthy, M. (2020). *A Systematic Exploration of Style Breach Detection Methods during Author Attribution Process* (Vol. 12). Special Issue. doi:10.5373/JARDCS/V12SP6/SP20201026

Quispe, L. V., Tohalino, J. A., & Amancio, D. R. (2013). *Using word embeddings to improve the discriminability of co-occurrence text networks*. arXiv preprint arXiv:2003.06279.

Reddy, P. & Mohan, T. & Raja, P. & Reddy, T. (2020). *A Novel Approach for Authorship Verification*. .10.1007/978-981-15-1097-7_37

Reddy, T., & Srilatha, M. (2020). Author Profiles Prediction Using Syntactic and Content-Based Features. doi:10.1007/978-981-15-1097-7_23

Research Note. (2020). Investigating the Viability of Stylometric Analysis to Attribute Authorship of Supreme Court Opinions. *Justice System Journal*. 10.1080/0098261X.2020.1723455

Ross, G. (2019). Tracking the evolution of literary style via Dirichlet–multinomial change point regression. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *183*. Advance online publication. doi:10.1111/rssa.12492

Rotari, Jander, & Rybicki. (2020). *A stylometric network analysis*. DOI: 10.1093/llc/fqz088

Sarwar, R., Porthaveepong, T., Rutherford, A., Rakthanmanon, T., & Nutanong, S. (2020). StyloThai: A Scalable Framework for Stylometric Authorship Identification of Thai Documents. *ACM Transactions on Asian and Low-Resource Language Information Processing.*, *19*, 1–15. doi:10.1145/3365832

Sarwar, R., Rutherford, A. T., Hassan, S. U., Rakthanmanon, T., & Nutanong, S. (2020, April 11). Native Language Identification of Fluent and Advanced Non-Native Writers. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(4), 1–9.

Schuster, T., Schuster, R., Shah, D., & Barzilay, R. (2020). The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*, 1–18. doi:10.1162/COLI_a_00380

Segarra, Eisen, Egan, & Ribeiro. (2020). A Response to Rosalind Barber's Critique of the Word Adjacency Method for Authorship Attribution. *ANQ: A Quarterly Journal of Short Articles, Notes and Reviews*, 1-6.

Sharma, H., Pundir, A., Yadav, N., Sharma, A., & Das, S. (Eds.), *Recent Trends in Communication and Intelligent Systems. Algorithms for Intelligent Systems*. Springer.

Suman, C., Bhattacharyya, P., & Chaudhari, R. (2020). Emoji Helps! A Multi-modal Siamese Architecture for Tweet User Verification. *Cognitive Computation*. Advance online publication. doi:10.1007/s12559-020-09715-7

Tarmizi, N., Saee, S., & Ibrahim, D. (2020). Author identification for Under-Resourced language (KadazanDusun). *Indonesian Journal of Electrical Engineering and Computer Science, 17*, 248-255. .v17.i1.pp248-25510.11591/ijeecs

Wang, Y., Garjami, J., Tsvetkova, M., Huu Hau, N., & Pho, K. H. (2012). Statistical approaches in literature: An application of principal component analysis and factor analysis to analyze the different arrangements about the Quran's Suras. *Digital Scholarship in the Humanities.*, *2020*(Mar), 10.

Zhu, H., Lei, L., & Craig, H. (2020). Prose, Verse and Authorship in Dream of the Red Chamber: A Stylometric Analysis. *Journal of Quantitative Linguistics*. Advance online publication. doi:10.1080/09296174.2020.1724677