The Key Factors and Their Influence in Authorship Attribution

Raheem Sarwar and Sarana Nutanong

City University of Hong Kong
Department of Computer Science
Hong Kong
rsarwar2-c@my.cityu.edu.hk, snutanong@cityu.edu.hk

Abstract. Authorship attribution has a long history started since 19th century. Existing studies have used different sets of stylometric features and computational methodologies on a variety of corpus with different lengths and genres. This study presents a protocol to perform a *systematic literature review (SLR)* to identify the best combination of stylometric features and computational methodology. Specifically, we formulate an SLR protocol that can be used to conduct a literature survey to help answer like (i) whether it is possible to identify the authorial style of an author regardless the genre and length of the text, and (ii) how to select specific stylometric features and computational methodology. We also conduct an example of how the proposed SLR protocol can be used as a template for publication extraction and filtering for an SLR on authorship attribution.

Keywords: Authorship attribution, Stylometric features, Computational methodologies

1 Introduction

Authorship Attribution (AA) problem is generally expressed as: given a disputed text and a set of candidate authors with their writing samples, find the author of the given disputed text from the set of candidate authors [1]. AA has a very long history started from 19th century and many approaches have been proposed for it. Existing approaches can be divided into two main tasks. Finding appropriate features of the language to quantify the writing style of authors, and forming efficient approaches to apply these features. A lot of stylometric features have been proposed so far including word lengths, sentence lengths, vocabulary richness and character frequencies. Rudman (1998) reported that almost 1 thousand measures has been proposed to quantify the writing styles of the authors [2]. During the last decade, this research areas has been extensively investigated by researchers in the fields of natural language processing [3,4], machine learning [5] and information retrieval [6].

Existing studies of AA used different sets of stylometric features and computational methodologies on a variety of corpus with different length and genre of the text [1,2,4,7–30]. We have formulated following research questions to address in the systematic literature review (SLR):

pp. 139–150; rec. 2016-02-12; acc. 2016-03-06

- **RQ1:** Which combination of the “set of stylometric features” and computational methodology is best in terms of accuracy in AA, and reasoning?
- **RQ2:** Is it possible in AA to identify the authorial style of the author regardless of the genre and length of corpus and without selecting specific stylometric features and computational methodology, and reasoning?

This paper presents a Systematic Literature Review (SLR) protocol to address key research questions in Authorship Identification. SLR is used for identification, evaluation and interpretation of all available research to specific research questions. To the best of our knowledge, this is the first SLR protocol formulated to address the authorship identification problem. Note that the nature of contribution of our work is introducing the SLR protocol rather than the study initiated from the protocol. The nature of our investigation is similar to the SLR protocols proposed by these publications [31–35]. Writing SLR protocol is important before we start the detailed review because the thoroughness of the protocol will ensure that the process remains rigorous. Developing an SLR protocol is consider prerequisite for detailed literature review in an area. A detailed review will be conducted as future work. The resultant protocol obtained from this investigation can be used to help investigate the scope of primary studies in which empirical evidence “contradicts” or “supports” with our theoretical hypotheses and to help generate new hypotheses. Specifically, our SLR protocol provides (i) a systematic means to select related studies in order to reduce biases through a well defined and comprehensive methodology; (ii) the information about the influence of some phenomenon based on empirical methods and wide range of settings. A consistent SLR study also provides evidence that phenomenon is transferable and robust, otherwise, the sources of the variations can be explored [36]. The rest of the paper is organized as follows. Section 2 presents the literature review. Section 3 presents the formation of the SLR protocol to answer the proposed research questions. Section 4 presents the preliminary results of this study. Section 5 presents the conclusion and future work.

2 Literature Review

Authorship attribution has a very long history started since 19th century. The first attempt to identify the author based on the writing style was made by Mendenhall [37] in 1887 followed by Zipf [8] and Yule [7] in 1932 and 1939, respectively. Later on, this problem was solved by performing the Bayesian statistical analysis on the frequencies of common words e.g., ‘to’, ‘and’ etc by Mosteller and Wallace [9] in 1964. Subsequently, Holmes [38] formulated a feature set to quantify the writing styles of the authors which is also known as Stylometry. The study of stylometry is concerned with statistical analyses of variations in the author’s literary style (represented as a set of features), which remains relatively unchanged across different documents [10,38]. Thus far, a variety of stylometric features have been proposed for AA including average sentence length, average word length [13], vocabulary richness [14], frequencies of punctuation [13], word endings [15], character n-grams [39], word n-grams [40], parts of speech n-grams [15], the organization of words, vocabulary distributions and the number of occurrences of particular word [41]. During the last decade, this research areas has been extensively investigated by researchers in the fields of natural language processing [3,4], machine learning [5] and information retrieval [6]. There are many techniques from machine learning and artificial intelligence that have been used for AA. In earlier days, the Bayesian statistical analysis [9] was used for authorship attribution; the recent techniques which have been used for authorship attribution include support vector machines [39] neural network [13,42], radial basis function networks [19], decision trees [18], and nearest neighbor classification [1]. Moreover, the markov chains [43], principal component analysis [17] and compression based techniques have also been used for AA [16].

3 Systematic Literature Review Protocol

According to Kitchenham [36], a systematic literature review (SLR) have three steps: (i) planning a review, (ii) conducting the review, and (iii) reporting the review. This paper focuses only on the first step, planning of a review, i.e., formulating an SLR protocol to address the research questions with preliminary results. An SLR protocol explains the methodology to conduct a literature review. The protocol decrease researchers bias to a specific set of publications [31]. For instance, without a predefined protocol, there is a possibility that the selection of primary studies may be driven by the expectations of the researcher [44]. Figure 1 shows the steps of the SLR protocol. The first and most important step is the formation of research questions. The next step is concerned with defining the search strategy (research process) to retrieve the primary research studies by exploring different publisher sites and index engines. The third step provides a method of how to filter irrelevant and less important studies. The next step involves assessing the quality of the selected primary research studies. Finally, the data collection and synthesis are performed. The details of each step is discussed in the following sections.

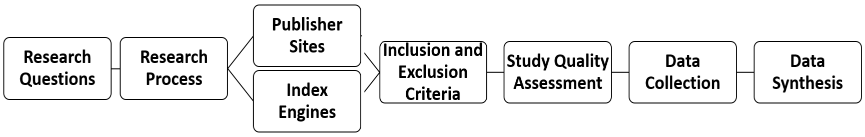

Fig. 1. Development of the Systematic Literature Review Protocol

3.1 Search Strategy

As explained in Section 1, we consider two research questions, RQ1 and RQ2.

We have developed the following strategy to formulate search queries to retrieve the primary studies to conduct the review.

1. **Derive Keywords:** Derive the main keywords from each research question.
2. **Derive Alternative Words:** Derive the alternative words or synonyms for each keyword obtained from research questions.
3. **Verification of Keywords:** Verify each keyword from the literature to ensure their correctness.
4. **Use Boolean Operators:** If bibliographic database provide the option, use Boolean "OR" operator to integrate alternative keywords and synonyms, and use the "AND" operator to integrate the major terms.

3.1.1 Results for 1 (Derive Keywords)

- **RQ1:** Stylometric Features, Computational Methodology, Authorship Attribution, Accuracy.
- **RQ2:** Authorship Attribution, Authorial Style, Author, Genre, Length, Corpus, Stylometric Features, Computational Methodology.

3.1.2 Results for 2 (Derive Alternative Keywords)

- **RQ1:**
- **Stylometric Features:**
- ("Stylometric Features" OR "authorial features" OR "stylometric properties" OR "stylometric analysis" OR "stylometric identification" OR "stylistic fingerprints" OR "linguistic fingerprint" OR "linguistic features ")
- **Computational Methodology:**
- ("Computational Methodology" OR "machine learning" OR "information retrieval" OR "bayesian statistical analysis" OR "support vector machines" OR "neural network" OR "radial basis function networks" OR "decision trees" OR "nearest neighbor classification" OR "markov chains" OR "principal component analysis" OR "compression based techniques" OR "latent dirichlet allocation" OR "feature transformation" OR "feature selection" OR "clustering" OR "supervised learning" OR "unsupervised learning" OR "semi supervised ensemble algorithm" OR "deep learning algorithm" OR "association rule" OR "instance based" OR "natural language processing" OR "statistical analysis")
- **Authorship Attribution:**
- ("Authorship Attribution" OR "author identification" OR "author recognition" OR "disputed authorship" OR "forensic authorship analysis" OR "author identity resolution" OR "stylometric identification")
- **Accuracy:**
- ("Accuracy" OR "enhance" OR "effective" OR "scalable" OR "experiment" OR "precision" OR "recall" OR "accurateness"
- **RQ2:**
- **Authorship Attribution:**

- (“Authorship Attribution” OR “author identification” OR “author recognition” OR “disputed authorship” OR “forensic authorship analysis” OR “author identity resolution” OR “stylometric identification”)
- **Authorial Style:**
- (“Authorial Style” OR “literary style”)
- **Author:**
- (“Author” OR “writer” OR “novelist” OR “biographer” OR “essayist” OR “dramatist” OR “playwright”)
- **Genre:**
- (“Genre” OR “type” OR “kind” OR “field” OR “email” OR “plays” OR “formal” OR “informal” OR “social Media”)
- **Length:**
- (“Length” OR “size” OR “short” OR “long” OR “chunk”)
- **Corpus:**
- (“Corpus” OR “text” OR “resource” OR “data”)
- **Stylometric Features:**
- (“Stylometric Features” OR “authorial features” OR “stylometric properties” OR “stylometric analysis” OR “stylometric identification” OR “stylistic fingerprints” OR “linguistic fingerprint” OR “linguistic features ”)
- **Computational Methodology:**
- (“Computational Methodology” OR “machine learning” OR “information retrieval” OR “bayesian statistical analysis” OR “support vector machines” OR “neural network” OR “radial basis function networks” OR “decision trees” OR “nearest neighbor classification” OR “markov chains” OR “principal component analysis” OR “compression based techniques” OR “latent dirichlet allocation” OR “feature transformation” OR “feature selection” OR “clustering” OR “supervised learning” OR “unsupervised learning” OR “semi supervised ensemble algorithm” OR “deep learning algorithm” OR “association rule” OR “instance based” OR “natural language processing” OR “statistical analysis”)

3.1.3 Results for 3 (Verification of Keywords)

- The correctness of all keywords from research questions have been verified from existing studies of Authorship Attribution.

3.1.4 Results for 4 (Use Boolean Operators)

- **RQ1:** (“Stylometric Features” OR “authorial features” OR “stylometric properties” OR “stylometric analysis” OR “stylometric identification” OR “stylistic fingerprints” OR “linguistic fingerprint” OR “linguistic features ”) AND (“Computational Methodology” OR “machine learning” OR “information retrieval” OR “bayesian statistical analysis” OR “support vector machines” OR “neural network” OR “radial basis function networks” OR “decision trees” OR “nearest neighbor classification” OR “markov chains” OR “principal component analysis” OR “compression based techniques”

OR “latent dirichlet allocation” OR “feature transformation” OR “feature selection” OR “clustering” OR “supervised learning” OR “unsupervised learning” OR “semi supervised ensemble algorithm” OR “deep learning algorithm” OR “association rule” OR “instance based” OR “natural language processing” OR “statistical analysis”) AND (“Authorship Attribution” OR “author identification” OR “authorship analysis” OR “author recognition” OR “disputed authorship” OR “authorship verification” OR “intrinsic plagiarism” OR “unconscious authorship” OR “obfuscate authorship” OR “marker of authorship” OR “analysis of authorship” OR “computational analysis of authorship” OR “linguistic pattern recognition” OR “forensic authorship analysis” OR “fighting authorship” OR “author identity resolution” OR “author profiling” OR “stylometric identification”) AND (“Accuracy” OR “enhance” OR “effective” OR “scalable” OR “experiment” OR “effect” OR “precision” OR “accurateness” OR “optimization” OR “robustness”) AND (“Reasoning” OR “causes” OR “basis” OR “root” OR “origin” OR “source”)

- **RQ2:** (“Authorship Attribution” OR “author identification” OR “authorship analysis” OR “author recognition” OR “disputed authorship” OR “authorship verification” OR “intrinsic plagiarism” OR “unconscious authorship” OR “obfuscate authorship” OR “marker of authorship” OR “analysis of authorship” OR “computational analysis of authorship” OR “linguistic pattern recognition” OR “forensic authorship analysis” OR “fighting authorship” OR “author identity resolution” OR “author profiling” OR “stylometric identification”) AND (“Authorial Style” OR “literary style” OR “authorial component”) AND (“Author” OR “writer” OR “novelist” OR “biographer” OR “essayist” OR “dramatist” OR “playwright”) AND (“Genre” OR “type” OR “kind” OR “field” OR “email” OR “plays” OR “formal” OR “informal” OR “social media” OR “contemporary American English”) AND (“Length” OR “size” OR “short” OR “long” OR “huge” OR “chunk”) AND (“Corpus” OR “text” OR “resource” OR “data”) AND (“Stylometric Features” OR “authorial features” OR “stylometric properties” OR “stylometric analysis” OR “stylometric identification” OR “stylistic fingerprints” OR “linguistic fingerprint” OR “linguistic features ”) AND (“Computational Methodology” OR “machine learning” OR “information retrieval” OR “bayesian statistical analysis” OR “support vector machines” OR “neural network” OR “radial basis function networks” OR “decision trees” OR “nearest neighbor classification” OR “markov chains” OR “principal component analysis” OR “compression based techniques” OR “latent dirichlet allocation” OR “feature transformation” OR “feature selection” OR “clustering” OR “supervised learning” OR “unsupervised learning” OR “semi supervised ensemble algorithm” OR “deep learning algorithm” OR “association rule” OR “instance based” OR “natural language processing” OR “statistical analysis”) AND (“Reasoning” OR “causes” OR “basis” OR “root” OR “origin” OR “source”)

3.2 Resources to be Searched

Different bibliographic databases are selected to extract relevant conference papers and journal articles. Bibliographic databases are chosen on the basis of research experience, preferences or suggested by other researchers and personal knowledge [45].

The resources utilized in this study are shown in Table 1.

Table 1. Resources to be searched

Publisher's Site	Index Engines
ACM Digital Library	Scopus
IEEE Xplore	Compendex
Wiley Inter Science	Google Scholar
Science Direct	Cite Seer
Springer Link	Inspec
Business Source Premier	ISI Web of Science

3.3 Documentation of Search Results

The documentation of the search results is important to make the query process precise and replicable [45]. During the systematic literature review, the following data of the retrieved publications will be recorded: Serial No, Bibliographic Database, Query Date, Search Strategy, Search String, Years, Number of publications retrieved, Initial Selection Decision, Final Selection Decision.

3.4 Publication Selection Criteria

Publication selection criteria is used to decide which research papers are included in, or excluded from, a systematic literature review. It helps to pilot the selection criteria for review on a subset of primary publications

3.4.1 Inclusion Criteria: The inclusion criteria used in this paper helps to determine which research paper should be considered for review. In this study only those articles, reports and research papers will be considered in which stylometric features are used for Authorship Attribution on the text of different genre and length. The inclusion criteria is as follows:

- Studies that use stylometric features Authorship Attribution.
- Studies that clearly describe the reasons of selecting a particular set of stylometric features and computational methodology.
- Studies that perform Authorship Attribution on the corpus of different length and genre.
- Studies that clearly describe the affect of length and genre of the text on the accuracy.

3.4.2 Exclusion Criteria: The following exclusion criteria is used to eliminate the irrelevant literature from selected research papers:

- Studies that does not focus on authorship attribution.
- Studies that are not written in English.
- Research work that does not highlight the affect of genre and length of the text and the selection of particular set of stylometric features and computational methodologies for Authorship Attribution.
- Primary literature will be reviewed on the basis of criteria mentioned in Table 2. The existing individual research papers contributing to a SLR is named as primary research; a SLR is the form of secondary study.

Table 2. Review Procedure of Primary Studies

Relevance Analysis Phase.	Inclusion and Exclusion Criteria
Uniqueness	Ensure the uniqueness of the publication. They must be written in English
Relevance	Read the title and abstract to ensure the relevance of the study with our research question, in case of ambiguity, go through introduction and conclusion of the publication
Full Text	Select the studies after reading full text

3.4.3 Publication Quality Assessment: The publication quality assessment (PQA) of selected papers will take place after applying relevance and selection criteria mentioned in Table 2. The PQA of the selected publications will be performed parallel to the phase of data extraction. For PQA the following research questions have been taken under consideration:

- Does the paper clearly describe the stylometric features and computational methodology adopted to perform Authorship Attribution as there are some studies which do not list the stylometric features adopted to conduct the study.
- Does the research paper clearly describe the reason to select the specific set of stylometric feature for a specific kind and length of text.
- Does the study compare the result with existing techniques.
- Is the researcher seems biased to mention positive results more than negative results?
  Each of the above point will be graded as “No” or “Yes” or “partial” or “N.A”.

3.4.4 Data Extraction: Data extraction is concerned with defining a procedure to get the relevant data from selected primary studies. A data extraction form is used to collect data from the selected studies to perform systematic literature review. Before the phase of data extraction, we will implement pilot data extraction. The review of selected primary studies will be undertaken by a single researcher who is responsible for data extraction. In case of an issue concerning the data extraction, a secondary reviewer will be approached for the guidance.

3.5 Data Synthesis

Data synthesis involves collecting and summarising the results of selected primary studies. The synthesis of extracted data can be categorized into five parts. The first part consists of stylometric features. The second part consists of the computational methodologies performed on these stylometric features. The third part is concerned with the effect of genre on stylometric features and computational methodologies. The fourth part is concerned with the effect of the length of the text on stylometric features and computational methodologies. The final part provides quantitative analyses on the results. The data for these five parts are synthesized and presented in the format similar to that of Table 3. In Table 3, the Frequency is the ratio of the primary studies which presents search area and the total number of selected primary studies. The percentage represents the percentage of the total primary studies in which the required information is clearly described.

Table 3. Data synthesis format. This table is only an example of format and intentionally does not present any specific figures.

Search Area	**Paper Title**	**Authors**	**Years**	**Frequency**	**Percentage**
Stylometric Features	Title 1	Authors	...	Freq. 1	...
	Title 1	Authors	...	Freq. 1	...
	.	.		.	
	.	.		.	
	.	.		.	
	Title n	Author n		Freq. n	
Computational Methodologies	...	...	...	...	...
Genre	...	...	...	...	...
Length	...	...	...	...	...
Reasoning	...	...	...	...	...

4 Preliminary Results

We are currently in the implementation phase of the SLR and we have got results for some of the aforementioned sections of the proposed protocol. These are Sections 3.2, 3.3 and 3.5. After applying the aforementioned search strategy mentioned in Section 3.2 on the specified bibliographic resources, we selected some primary studies from retrieved studies based on the inclusion/exclusion criteria mentioned in 3.5. The preliminary results of this study based on existing studies can be summarised as follows:

- **RQ1:**The selection of the stylometric features affects the accuracy of the authorship attribution (AA). Moreover, the selection of appropriate computational methodology for a specific set of stylometric feature increases the accuracy of the AA [1, 2, 7–19].
- **RQ2:** The text of different genres require different set of stylometric features to obtain satisfactory results [4, 20, 21]. Genre-dependent stylometric features outperform the genre-independent stylometric features [22, 27, 30]. The accuracy of AA task is highly dependent on the length of the text, long text produce satisfactory results as compared to short text with same set of stylometric features, however, satisfactory results can be obtained with short text by selecting appropriate stylometric features [7, 26].

The preliminary results mentioned above will help to answer our proposed research question in the SRL.

5 Conclusions and Future Work

In this study, we propose a systematic literature review (SLR) protocol to identify the key factors and their influence in the field of authorship attribution. Our proposed SLR protocol can be used to create a well-defined literature survey in the area of Authorship Attribution. Specifically, we focus on exploring different stylometric feature sets and computational methodologies that can be adopted to increase the accuracy of Authorship Attribution. Our protocol can be used to define the scope of primary studies in which empirical evidence “contradicts” with or “supports” our theoretical hypotheses and will help to generate new hypotheses. As future work, we plan to apply our proposed protocol to conduct a comprehensive SLR study on our proposed research questions.

References

1. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for information Science and Technology **60** (2009) 9–26
2. Rudman, J.: The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities **31** (1997) 351–365
3. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. Computers and the Humanities **35** (2001) 193–214
4. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational linguistics **26** (2000) 471–495
5. Khosmood, F., Levinson, R.: Toward unification of source attribution processes and techniques. In: Machine Learning and Cybernetics, 2006 International Conference on, IEEE (2006) 4551–4556

6. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. In: Information Retrieval Technology. Springer (2005) 174–189
7. Yule, G.U.: On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. Biometrika **30** (1939) 363–390
8. Zipf, G.K.: Selected studies of the principle of relative frequency in language. (1932)
9. Mosteller, F., Wallace, D.: Inference and disputed authorship: The federalist. (1964)
10. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and linguistic computing **13** (1998) 111–117
11. Argamon, S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. In: ACH/ALLC. (2005)
12. Argamon, S.: Interpreting burrows's delta: geometric and probabilistic foundations. Literary and Linguistic Computing **23** (2008) 131–147
13. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering **11** (2005) 397–415
14. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proceedings of the twenty-first international conference on Machine learning, ACM (2004) 62
15. Madigan, D., Genkin, A., Lewis, D.D., Argamon, S., Fradkin, D., Ye, L.: Author identification on the large scale. In: Proc. of the Meeting of the Classification Society of North America. (2005) 13
16. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Physical Review Letters **88** (2002) 048702
17. Juola, P., Baayen, R.H.: A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguistic Computing **20** (2005) 59–67
18. Uzuner, Ö., Katz, B.: A comparative study of language models for book and author recognition. In: Natural Language Processing–IJCNLP 2005. Springer (2005) 969–980
19. Pandian, A., Sadiq, M.A.K.: Authorship categorization in email investigations using fisher's linear discriminant method with radial basis function. Journal of Computer Science **10** (2014) 1003
20. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Text genre detection using common word frequencies. In: Proceedings of the 18th conference on Computational linguistics-Volume 2, Association for Computational Linguistics (2000) 808–814
21. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN- **23** (2003) 321–346
22. Amasyalı, M.F., Diri, B.: Automatic turkish text categorization in terms of author, genre and gender. In: Natural Language Processing and Information Systems. Springer (2006) 221–226
23. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and linguistic Computing (2010) fqq013
24. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing **22** (2007) 405–417
25. Mikros, G.K., Argiri, E.K.: Investigating topic influence in authorship attribution. In: PAN. (2007)
26. Eder, M.: Does size matter? authorship attribution, small samples, big problem. Proceedings of Digital Humanities (2010) 132–135
27. Rybicki, J., Eder, M.: Deeper delta across genres and languages: do we really need the most frequent words? Literary and Linguistic Computing (2011) fqr031

28. Jockers, M.L., Witten, D.M.: A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing (2010) fqq001
29. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Artificial Intelligence: Methodology, Systems, and Applications. Springer (2006) 77–86
30. Kelih, E., Antić, G., Grzybek, P., Stadlober, E.: Classification of author and/or genre? the impact of word length. In: Classification—the Ubiquitous Challenge. Springer (2005) 498–505
31. Khan, A., Basri, S., Amin, F., Teknologi, U., Perak, T., Studies, I.: Communication risks and best practices in global software development during requirements change management: A systematic literature review protocol. Research Journal of Applied Sciences, Engineering and Technology **6** (2013) 3514
32. Rehman, S., Khan, S.U.: Swot analysis of software quality metrics for global software development: A systematic literature review protocol. IOSR Journal of Computer Engineering **2** (2012)
33. Khan, S.U., Niazi, M., Ikram, N.: Software development outsourcing relationships trust: a systematic literature review protocol. Evaluation and Assessment in Software Engineering, EASE (2010)
34. Alam, A.U., Khan, S.U.: Knowledge sharing management in offshore software development outsourcing relationships from vendors' perspective: A systematic literature review protocol. In: Software Engineering (MySEC), 2011 5th Malaysian Conference in, IEEE (2011) 469–474
35. Qureshi, N., Ikram, N., Bano, M., Usman, M.: Empirical evidence in software architecture: a systematic literature review protocol. In: The Sixth International Conference on Software Engineering Advances. (2011) 534–538
36. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele University **33** (2004) 1–26
37. Mendenhall, T.C.: The characteristic curves of composition. Science (1887) 237–249
38. Holmes, D.I.: Authorship attribution. Computers and the Humanities **28** (1994) 87–106
39. Hedegaard, S., Simonsen, J.G.: Lost in translation: Authorship attribution using frame semantics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics (2011) 65–70
40. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics (2011) 288–298
41. Juola, P.: Authorship attribution for electronic documents. In: Advances in digital forensics II. Springer (2006) 119–130
42. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology **57** (2006) 378–393
43. Baayen, H., van Halteren, H., Neijt, A., Tweedie, F.: An experiment in authorship attribution. In: 6th JADT. (2002) 29–37
44. Khan, A.A., Basri, S., Dominic, P.: Communication risks in gsd during rcm: Results from slr. In: Computer and Information Sciences (ICCOINS), 2014 International Conference on, IEEE (2014) 1–6
45. Chen, L., Ali Babar, M., Zhang, H.: Towards an evidence-based understanding of electronic data sources. (2010)