Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 2, 432-440 (2020)
www.astesj.com
ASTESJ
ISSN: 2415-6698

Author Identification for Marathi Language

C. Namrata Mahender*, Ramesh Ram Naik, Maheshkumar Bhujangrao Landge

Department of CS and IT, Dr.B.A.M.University, Aurangabad-431004 (MS), India

ARTICLE INFO

Article history:
Received: 22 July, 2019
Accepted: 09 December, 2019
Online: 04 April, 2020

Keywords:
Plagiarism detection
Author identification
Marathi language.

ABSTRACT

This is era of new technology; most of information is collected from internet, web sites. Some people uses data from research papers, thesis, and website as it is and publish as their own research without giving proper acknowledgement. This term is known as plagiarism. There are two types of plagiarism detection methods, i) Extrinsic plagiarism detection ii) Intrinsic plagiarism detection. Through extrinsic plagiarism utilizing reference corpus plagiarism is observed, while in intrinsic plagiarism identification, using author's writing style, plagiarism can be identified. If the anonymous text is written by unknown author. By using authorship analysis we can find original author of text. Authorship analysis is having three types i)Author identification ii) Author characterization and iii) Similarity detection. This paper mainly focuses on author identification for Marathi language. To calculate projection in two different files, we used feature vectors of main author file and summary file of other authors. The result of average projection shows, there is similarity in main author file and summary file of different authors, it also shows summary file of each author is having impact of main author file.

1. Introduction

Plagiarism includes copying material, every word from phrase or as a paraphrase, from any book to websites, course notes, oral or visual displays, lab reports, pc assignments, or artistic works. Plagiarism includes reproducing any individual else's work, whether or not it be posted article, chapter of a book, a paper from a buddy or some file, or whatever. In addition, plagiarism involves the exercise of employing another person to alter or revise the work that a student submits as his or her own, whoever that other man or woman may be. Authorship identification is the ability to identify unidentified authors based on their previous work and statements. The main method in authorship identification is to look at and identify features by an author using stylometric features. We can find the writing style of author by identifying textual features that they used while writing document [1].

1.1. Authorship Analysis

Authorship analysis is a method of analyzing the features of the writing part in order to draw conclusions from its authorship [1]. Authorship analysis having three types: i) Authorship Identification, ii) Authorship characterization, iii) Similarity detection.

A. Authorship identification: It defines the likelihood of a part of the writing being produced by a specific author by examining the author's other writings.

B. Authorship characterization: Authorship characterization reviews the character-istics of an author and produces the author profile based on his or her writing.

C. Similarity detection: Similarity detection examines several pieces of writing and judges whether they have been published by a single author without actually identifying the author [1].

2. Literature Survey

The PAN workshop brought together experts and researchers around the exciting and future-oriented topics of plagiarism detection, authorship identification, and the detection of social software misuse. It started in 2009. But relevant to Plagiarism the track started in 2011. The table1 shows that PAN Features used, and technique applied from the year 2011 to 2018.

*C. Namrata Mahender, Department of CS and IT, Dr.B.A.M.University, Aurangabad-431004 (MS), INDIA, nam.mah@gmail.com

https://dx.doi.org/10.25046/aj050256

Table 1: PAN Features and technique used from the year 2011 to 2018.

Reference Number	Features	Technique used
[2]	Bag of words features are used	In this paper author used Approach over known authors documents, using support vector machines. author treat each paragraph as a separate document and apply the n-cut clustering algorithm
[3]	1. Lexical features 2. Character level 3.various length-related features 4. syntax related features	In this paper author was used Support vector machine classifier for classification.
[4]	Language-dependent Content and Stylometric Features	Author used SVM and random forests as classifiers and regressors.
[5]	Word ngrams, Character ngrams, POS ,tag ngrams, Word lengths, Sentence lengths ,Sentence length ngrams ,Word richness ,Punctuation ngrams ,Text shape ngrams.	Author explored three different regressor algorithms: trees, random forests, and support vector machines.
[6]	n-gram	PPM (Prediction by Partial Matching) compression algorithm based on an n-gram statistical model.
[7]	phrase-level and lexical-syntactic features 1. Word prefixes 2. Word sufixes 3. Stopwords 4. Punctuation marks 5. N-grams(one gram to Fivegram features calculated) 6. Skip-grams (one gram to Fivegram features calculated) 7. Vowel combination 8. Vowel permutation	A similarity vector using the LSA algorithm for each word in the test documents Different distance/similarity measures were tested, including the Jaccard similarity for the vocabulary feature vector, the cosine similarity for the Frequency vector of all the combined Lexical syntactic features and Chebyshev Distance, Euclidean distance and cosine similarity for the LSA vectors.
[8]	1. Character 2. Words 3. Lemma and Part of Speech	Our method is based on the analysis of the average similarity (ASUnk) of an unknown authorship text with the closeness to each of the samples of an author, comparing it to the Average Group Similarity (AGS) between samples of an author.
[9]	Bag of words using character n-grams	Author used Ensemble Particle Swarm Model Selection (EPSMS) for the selection of classification models for each data set. For classification we used the neural network classifier implemented in the CLOP toolbox
[10]	stylometric features 1. Basic features 2.Lexical features 3. Character features 4. Syntactic features 5.Coherence features	Author follows the unmasking approach.
[11]	1.length of the sentences, 2.variety of vocabulary, 3. Words, n-characters grams, n-4. Words gram, punctuation marks.	Author compares all documents inside a corpus using the cosine similarity, euclidean distance or the correlation coefficient. For the task of Author Verification, we used the Classification and Regression Trees (CART) algorithm which constructs binary trees using the features and thresholds that

		yield the largest information gain at each node
[12]	profiles of character 3-grams for representing information about the Different categories of authors.	Baseline (accuracy) obtained in cross-genre classification by age and gender using Naive Bayes, tf-idf word representation.
[13]	word bag, stop word bag, punctuation bag, part of speech (POS) bag	KNN Algorithm is used
[14]	1. counting text elements 2. constructing syntactic n-grams	Integrated syntactic graph is used.
[15]	1.Char Sequences 2.Word Uni-grams 3. POS-tags Features	PCA Linear SVC
[16]	phoneme-based features, character-based features, token-based features, syntax-based features, semantic-based features	k-NN classifier
[17]	signatures, chat slang, context, emotionality, semantic similarity, Jaccard similarity and BOW	NB classifier
[18]	Stylistic Features 1.Stylometry based approaches 2.Content based approaches 3.Topic based approaches	Navies Bayes, Support Vector Machine, Random Forest, J48 and Logistics. These algorithms was used.
[19]	lexical, syntactic and graph-based features	Support Vector Machines (SVM).
[20]	character n-grams	Vector Space Model, Similarity Overlap Metric
[21]	Basic Statistics, Token Statistics, Grammar Statistics, Stop-Word Terms, Pronoun Terms, Slang Terms, Intro-Outro Terms, Bigram Terms, Unigram Terms, and Terms.	Supervised vote/veto meta-classifier approach
[22]	Stylometric features or word n-grams.	k-NN classifier
[23]	n-grams	Distance measure technique used.
[24]	n-Grams	Support Vector Machine classifier
[25]	n-grams	Local n-gram Technique is used.
[26]	Bag of words, Bigram, Trigram, Comma Dots, Numbers, Capitals, Words per paragraph, Sentences per paragraph, Square brackets.	Support Vector Regression and Neuronal Networks models
[27]	n-grams of POS tag sequences	vector space model
[28]	stylistic and statistical features	SVM, Bayes, KNN
[29]	stylometric features ranging from characters to syntactic and semantic units	SVM
[30]	n-grams	SVM
[31]	First words of sentences or lines, nouns, verbs, punctuation.	principal component analysis
[32]	stylometric properties, grammatical characteristics and pure statistical features	SVM classifier
[33]	Linguistic Features	SVM
[34]	n-grams	LSA
[35]	Unigram-Tf-idf, Unigram Character, Character4-gram	GenIM method
[36]	Stylistic Total number of words Average number of words per sentence	SVM, K-means clustering Algorithm implemented in CLUTO

	Binary feature indicating use of quotations Binary feature indicating use of signature Percentage of all caps words Percentage of non-alphanumeric characters Percentage of sentence initial words with first letter capitalized Percentage of digits Number of new lines in the text Average number of punctuations (!?.;:,) per sentence Percentage of contractions (won't, can't) Percentage of two or more consecutive non-alphanumeric characters. Lexical Bag of words (freq. of unigrams) Perplexity Perplexity values from character 3-grams Syntactic Part-of-Speech (POS) tags Dependency relations Chunks (unigram freq.)	
[37]	Elimination of stopwords, punctuation symbols and xml tags	Rocchio, Naïve Bayes and Greedy

3. Text Corpus

Similar to other language work, work in the Marathi language is also appreciable. But the work is not accessible as an online resource, so far it's offline. Actually, there is no generic Marathi text corpus accessible. For the development of text corpus, we have considered 10 paragraphs for taking summary from 50 users in their own writing. We have used 500 summary files from 50 users as a database for author identification.

रात्रीचे जेवण लवकर घ्यावे.त्यामुळे त्याचे पचनही चांगले होते आणि स्थूलपणा कमी होण्यास मदत होते.प्रत्येकाने ही तत्त्वे नेटाने,नित्यनेमाने पाळली तर स्थूलपणा कमी होण्यास मदत होईल आणि चरबीचे थर निघून जातील.तसेच शरीर हलके होईल.त्यामुळे मन प्रसन्न,आनंदी राहील.स्थूलपणा हा बदललेल्या जीवनशैलीचा दुष्परिणाम आहे.घरी बनवलेले रूचकर,सात्त्विक जेवण ज्यात वरण,भात,चपाती वा भाकरी,भाजी,कोशिबीर,दही,उसळ या सर्वाचा समावेश असेल,तर स्थूलपणाला सहज रामराम ठोकणे शक्य आहे.

Figure 1: Sample file from database

अन्न हे पूर्णब्रहम आहे अन्न हे रुचकर करण्यासाठी त्यात आपले मनशांती असणे गरजेचे आहे जेंव्हा आपण जेवणात अन्न ग्रहण करतो तेंव्हा आपण व्यवस्थित चावणे हे गरजेचे असते.जेवणात वरण,भात,चपाती,वा भाकरी, उसळ,कोशिंबीर असणे हे स्थूलपनाला रामराम ठोकणे आहे पण हे अन्न आपण नित्यनेमाने करणे आवश्यक आहे.

Figure 2: Sample Summary written by Author

4. Proposed System

We would like to propose a system for Author Identification in Marathi Language. The system workflow is given below:

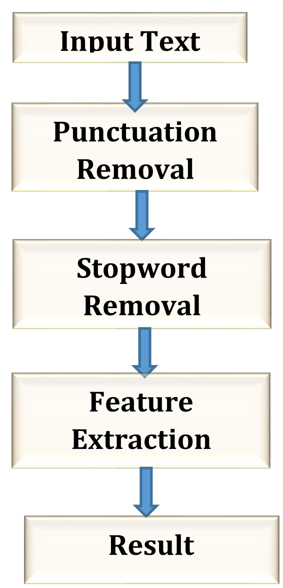

Figure3: Proposed System for Author Identification for Marathi Language

4.1. Input Text

First the system reads two files. Main file and summary of written by Authors file. The file format is .txt

4.2. Punctuation removal

This step removes the punctuations present in the file, e.g. punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

4.3. Stopword Removal

Stop words are simply a set of words widely used in any language. Here are the Stopwords:

Table 2. List of Stopwords

या	त्यांनी	हा	पण
व	सुरु	ही	जेंव्हा
यांनी	करून	करण्यात	त्या
हे	जर	याच्या	त्याच्या
तर	असून	ता	मात्र
ते	आले	तेंव्हा	परंतु
असे	त्यामुळे	हा	पण

Table 3: Features of Original Sample files

main files	avg sen len by char	avg sen len by word	hapax legema	hapax dislegama	avg word freq class	avg sen len
OG_File1	1198	57	423.41	0.11	1.79	7
OG_File2	1441	74	441.88	0.19	1.55	9
OG_File3	1612	79	443.08	0.1	1.77	9
OG_File4	2797	128	492.72	0.07	1.84	7
OG_File5	2896	154	508.75	0.09	1.95	7
OG_File6	2757	141	499.04	0.06	1.89	7
OG_File7	2841	141	503.69	0.04	1.82	7
OG_File8	991	63	417.43	0.12	1.69	13
OG_File9	740	30	358.35	0	1	4
OG_File10	1173	44	417.43	0.1	1.76	11

5. Feature Extraction

Feature extraction can be defined as the process of extracting a set of new features from the set of features generated in the selection stage feature. Feature extraction is a basic and fundamental step to pattern Recognition and machine learning problem. There is no text corpus available for Marathi language.

We concentrated on two major features: Lexical features and Vocabulary richness features. These include features like Average sentence length by word, Average sentence length by character ,AvgWordFrequencyClass, Avg sentence length, Hapax legomenon, Hapax dislegemena.

We have extracted the following features:

5.1. Lexical features

1. Average length of sentence by word
2. Average length of sentence by character
3. AvgWordFrequencyClass
4. Avg sentence length

5.2. Vocabulary richness features

1. Hapax legomenon
2. Hapax dislegemena

Hapax Legomena and Hapax DisLegemena

Hapax Legomena is a term that appears only once in a sense, either in the written record of the whole language, a single text. Hapax legomenon it is a Greek phrase which is means something that told onetime only.

Similarly, Hapax DisLegemena is the word that is used twice. Following table3 shows that features of original sample files from database.

Table 4: Features of Author1 files

Files ↓	Avg_SentLenghtByCh	Avg_SentLenghtByWord	hapaxLegemena	hapaxDisLegemena	Avg Word Frequency Class	Avg sentence length
File1	758.0	44.0	391.20	0.054	1.7	15
File2	1049.0	68.0	426.26	0.24	1.53	34
File3	943.0	57.0	409.43	0.183	1.65	14
File4	1149.0	67.0	423.41	0.084	1.75	17
File5	1243.0	75.0	436.94	0.072	1.78	15
File6	1465.0	90.0	453.25	0.22	1.52	45
File7	754.0	44.0	395.12	0.04	1.92	15
File8	572.0	41.0	376.12	0.131	1.76	14
File9	538.0	25.0	349.65	0.064	1.87	8
File10	645.0	28.0	361.09	0.0 0	1.0	14

Table 5: Features of Author2 files

Files ↓	Avg_SentLenghtByCh	Avg_SentLenghtByWord	hapaxLegemena	hapaxDisLegemena	Avg Word FrequencyClass	Avg sentence length
File1	877.0	49.0	397.02	0.1041	1.81	12
File2	1076	59.0	411.08	0.113	1.75	10
File3	1296.0	71.0	429.04	0.089	1.83	18
File4	1366.0	72.0	434.38	0.069	1.87	15
File5	1103	84.0	438.35	0.059	1.82	14
File6	678	82.0	538.0	0.079	1.79	16
File7	899	65.0	458.0	0.085	1.84	15
File8	523.0	30.0	349.65	0.033	1.84	8
File9	442.0	19.0	317.80	0.0	1.0	5
File10	869.0	37.0	380.66	0.04	1.84	9

Table 6: Features of Author3 file

Files ↓	Avg_SentLenghtByCh	Avg_SentLenghtByWord	hapaxLegemena	hapaxDisLegemena	AvgWordFrequencyClass	Avg sentence length
File1	777.0	47.0	395.12	0.1063	1.80	23
File2	880	67.0	412.11	0.13	1.82	20
File3	1390.0	86.0	449.98	0.154	1.87	29
File4	1230	82.0	468.25	0.123	1.85	22
File5	1178	86	434.0	0.14	1.78	24
File6	879	81.0	398.0	0.13	1.87	22
File7	758	58.0	369.0	0.15	1.83	20
File8	627.0	41.0	376.12	0.176	1.62	14
File9	598.0	34.0	361.09	0.23	1.62	11
File10	686.0	36.0	371.35	0.051	1.90	36

Table 7: Features of Author4 file

Files ↓	Avg_SentLenghtByCh	Avg_SentLenghtByWord	hapaxLegemena	hapaxDisLegemena	AvgWordFrequencyClass	Avg sentence length
File1	758.0	47.0	389.18	0.05 0	1.71	23
File2	796	49.0	387.10	0.02	1.74	22
File3	947.0	51.0	397.02	0.02	1.88	25
File4	864.0	53.0	434.0	0.03	1.85	23
File5	1164	52.0	489	0.086	1.83	20
File6	1516.0	84.0	0.051	445.43	1.82	10
File7	1526.0	94.0	456.43	0.1392	1.67	19
File8	496.0	29.0	343.39	0.074	1.77	14
File9	565.0	27.0	343.39	0.0	1.0	13
File10	1071.0	53.0	404.30	0.058	1.82	18

Table 8: Features of Author5 file

Files ↓	Avg_SentLenghtByCh	Avg_SentLenghtByWord	hapaxLegemena	hapaxDisLegemena	AvgWordFrequencyClass	Avg sentence length
File1	794.0	45.0	391.20	0.090	1.78	11
File2	1056.0	64.0	418.96	0.157	1.72	16
File3	1020.0	56.0	398.21	0.18	1.85	14
File4	2093.0	104.0	468.21	0.061	1.83	9
File5	1524.0	102.0	485.11	0.071	1.84	10
File6	1754.0	107.0	480.12	0.078	1.86	12
File7	1825.0	111.0	475.35	0.11	1.74	16
File8	715.0	46.0	387.12	0.12	1.72	23
File9	631.0	31.0	358.35	0.0	1.0	10
File10	812.0	31.0	378.41	0.07	1.86	10

6. Result

$$\textbf{projection} = \frac{\overrightarrow{AS}.\overrightarrow{OS}}{|\overrightarrow{AS}.\overrightarrow{OS}|} \quad (1)$$

$\overrightarrow{AS}$ Feature vector of summary file written by author

$\overrightarrow{OS}$-> Feature vector of main author file from database

Table 9: Projections of main author file on summary file written by author

Projection of File1			Projection of File2			Projection of File3		
Feature vector of original file	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**
O1 S1	A1 S1	1259.96	O2 S2	A1 S2	1502.67	O3 S3	A1 S3	1656.90
O1 S1	A2 S1	1267.24	O2 S2	A2 S2	1505.64	O3 S3	A2 S3	1671.39
O1 S1	A3 S1	1260.77	O2 S2	A3 S2	1493.81	O3 S3	A3 S3	1671.71
O1 S1	A4 S1	1260.08	O2 S2	A4 S2	1490.71	O3 S3	A4 S3	1659.55
O1 S1	A5 S1	1263.03	O2 S2	A5 S2	1504.15	O3 S3	A5 S3	1664.60
Projection of File4			**Projection of File5**			**Projection of File6**		
Feature vector of original file	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**
O4 S4	A1 S4	2797.49	O5 S5	A1 S5	2904.78	O6 S6	A1 S6	2783.81
O4 S4	A2 S4	2817.58	O5 S5	A2 S5	2882.72	O6 S6	A2 S6	2471.87
O4 S4	A3 S4	2791.57	O5 S5	A3 S5	2896.68	O6 S6	A3 S6	2719.10
O4 S4	A4 S4	2722.88	O5 S5	A4 S5	2870.66	O6 S6	A4 S6	2789.41
O4 S4	A5 S4	2839.97	O5 S5	A5 S5	2917.76	O6 S6	A5 S6	2794.38
Projection of File7			**Projection of File8**			**Projection of File9**		
Feature vector of original file	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**	**Feature vector of original file**	**Feature Vector of Author file**	**Projection**
O7 S7	A1 S7	2753.51	O8 S8	A1 S8	1059.29	O9 S9	A1 S9	816.28
O7 S7	A2 S7	2763.22	O8 S8	A2 S8	1057.72	O9 S9	A2 S9	810.570
O7 S7	A3 S7	2777.38	O8 S8	A3 S8	1066.46	O9 S9	A3 S9	819.15
O7 S7	A4 S7	2869.40	O8 S8	A4 S8	1054.22	O9 S9	A4 S9	818.94
O7 S7	A5 S7	2879.50	O8 S8	A5 S8	1072.00	O9 S9	A5 S9	820.94
Projection of File10								
Feature vector of original file	**Feature Vector of Author file**	**Projection**						
O10 S10	A1 S10	1228.20						
O10 S10	A2 S10	1242.74						
O10 S10	A3 S10	1230.19						
O10 S10	A4 S10	1245.55						
O10 S10	A5 S10	1240.36						

Table 10: Average projection of main author on dependent author

Name of Projection Files	Average projection of each file
File1	1262.22
File2	1499.401
File3	1664.835
File4	2793.904
File5	2894.525
File6	2711.718
File7	2808.606
File8	1061.944
File9	817.1817
File10	1237.416

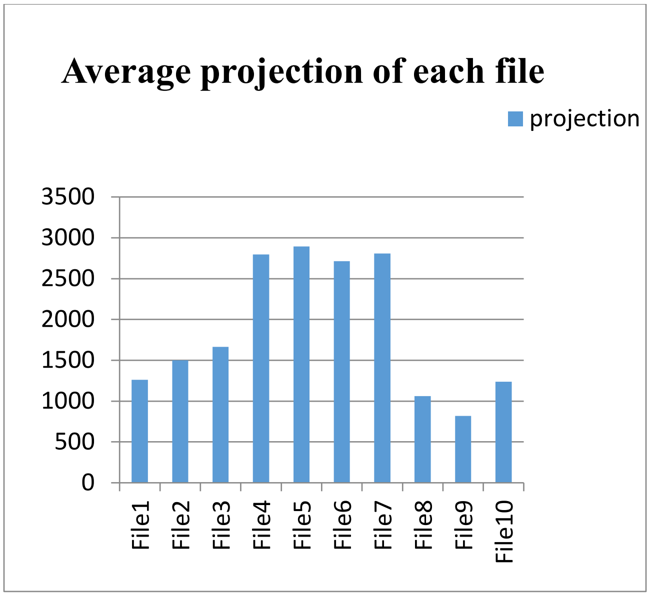

Figure 4: Average projection of each file

Above figure 4 shows average projection of 10 files. We have calculated feature vector of main author file and feature vector of summary file written by author, we calculated projection these two vectors for 10 different sample summary files of five authors. It shows there is similarity in main author file and summary file of each author. Summary file of author is having impact of main author file. Above graph shows file number 4,5,6,7 are having more projection of main author file.

7. Conclusion

Authorship identification is the ability to identify unidentified authors based on their previous work and statements. We have created database of 500 summary files from 50 users for author identification. After doing literature survey on features used for author identification, we selected some features like Lexical features and vocabulary richness features. By using feature vector of main author file and summary file of authors, we calculated projection of 10 files. The result of average projection shows, there is similarity in main author file and summary file of different authors. The figure4 shows summary file of each author is having impact of main author file, Summary file number 4,5,6,7 are having more projection of main author file. Currently, most of Marathi native speakers are contributing their research for various topics in Marathi language, but some of researchers are using information from various sources like research papers, books, thesis without giving acknowledgement. There is need to restrict these type of conditions. There is no Author identification tool available for Marathi language. This tool will be helpful to perform quality research in Marathi language.

Acknowledgment

Authors would like to acknowledge and thanks to CSRI DST Major Project sanctioned No.SR/CSRI/71/2015(G), Computational and Psycholinguistic Research Lab Facility supporting to this work and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.

References

[1] 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(3), 378-393. 2006, DOI: 10.1002/asi.20316.
[2] Akiva, Navot. “Authorship and Plagiarism Detection Using Binary BOW Features” CLEF (Online Working Notes/Labs/Workshop) 2012.
[3] Argamon, S., & Juola, P. “Overview of the international authorship identification competition at PAN-2011”. In CLEF (Notebook Papers/Labs/Workshop). 2011.
[4] Bartoli, A., De Lorenzo, A., Laderchi, A., Medvet, E., & Tarlao, F. “An author profiling approach based on language-dependent content and stylometric features”. In Conference and Labs of the Evaluation forum (Vol. 1391). CEUR. 2015.
[5] Bartoli, A., Dagri, A., De Lorenzo, A., Medvet, E., & Tarlao, F. “An author verification approach based on differential features”. In Conference and Labs of the Evaluation forum (Vol. 1391). CEUR. 2015.
[6] Bobicev, V. “Authorship detection with PPM”. In Proceedings of CLEF. 2013.
[7] Alhijawi, B., Hriez, S., & Awajan, A. “Text-based Authorship Identification-A survey”. In 2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT) (pp. 1-7). IEEE.2018. DOI: 10.1109/ISIICT.2018.8613287
[8] Castro, D., Adame, Y., Pelaez, M., & Muñoz, R. “Authorship verification, combining linguistic features and different similarity functions”, CLEF (Working Notes). 2015.
[9] Escalante, H. J. “EPSMS and the Document Occurrence Representation for Authorship Identification”, Notebook for PAN at CLEF 2011.
[10] Feng, V. W., & Hirst, G. “Authorship verification with entity coherence and other rich linguistic features”, In Proceedings of CLEF (Vol. 13) 2013.
[11] Fréry, J., Largeron, C., & Juganaru-Mathieu, M. “Ujm at clef in author identification”, Proceedings CLEF-2014, Working Notes, 1042-1048. 2014.
[12] Ucelay, M. J. G., Villegas, M. P., Funez, D. G., Cagnina, L. C., Errecalde, M. L., Ramírez-de-la-Rosa, G., & Villatoro-Tello, E. “Profile-based Approach for Age and Gender Identification”, In CLEF (Working Notes) (pp. 864-873). 2016.
[13] MR Ghaeini. “Intrinsic Author Identification Using Modified Weighted KNN” -Notebook for PAN at CLEF 2013.
[14] Gómez-Adorno, H., Sidorov, G., Pinto, D., & Markov, I. “A graph based authorship identification approach”. Working notes papers of the CLEF, 2015.
[15] HaCohen-Kerner, Y., Miller, D., Yigal, Y., & Shayovitz, E. “Cross-domain Authorship Attribution: Author Identification using char sequences, word unigrams, and POS-tags features”, Working Notes of CLEF. 2018.
[16] Halvani, O., Steinebach, M., & Zimmermann, R. “ Authorship verification via k-nearest neighbor estimation”. Notebook PAN at CLEF. 2013.
[17] Hernández, D. I., Guzmán-Cabrera, R., & Reyes, “A. Semantic-based Features for Author Profiling Identification” First insights Notebook for PAN at CLEF 2013.

[18] Pervaz, I., Ameer, I., Sittar, A., & Nawab, R. M. A. "Identification of Author Personality Traits using Stylistic Features", Notebook for PAN at CLEF 2015.
[19] Vilariño, D., Pinto, D., Gómez, H., León, S., & Castillo, E. "Lexical-syntactic and graph-based features for authorship verification". In Proceedings of CLEF, 2013.
[20] Jayapal, A., & Goswami, B. "Vector space model and Overlap metric for Author Identification", Notebook for PAN at CLEF 2013.
[21] Kern, R., Klampfl, S., & Zechner, M. "Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification", Notebook for PAN at CLEF 2012.
[22] Kern, R. "Grammar Checker Features for Author Identification and Author Profiling" In CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers.2013.
[23] Kocher, M., & Savoy, J. "Author Identification" Working Notes Papers of the CLEF.2015.
[24] Kourtis, Ioannis, and Efstathios Stamatatos "Author identification using semi- supervised learning", CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, the Netherlands. 2011.
[25] Layton, Robert, Paul Watters, and Richard Dazeley" , Local n-grams for Autho Identification". Notebook for PAN at CLEF 2013.
[26] Ledesma, P., Fuentes, G., Jasso, G., Toledo, A., & Meza, I. "Distance learning for author verification" In Proceedings of the conference pacific association for computational linguistics, PACLING (Vol. 3, pp. 255-264). 2003.
[27] López-Anguita, Rocío, Arturo Montejo-Ráez, and Manuel Carlos Díaz-Galiano "Complexity Measures and POS n-grams for Author Identification in Several Languages", SINAI at PAN@ CLEF 2018.
[28] Mechti, S., Jaoua, M., Faiz, R., & Belguith, L. H. "On the Empirical Evaluation of Author Identification Hybrid Method".2015.
[29] Mikros, G. K., & Perifanos, K. "Authorship identification in large email collections: Experiments using features that belong to different linguistic levels". Notebook for PAN at CLEF, 2011.
[30] Moreau, E., Jayapal, A., & Vogel, C. "Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm", Notebook for PAN at CLEF In Working Notes for CLEF 2014 Conference (Vol. 1180, p. 12). CEUR Workshop Proceedings. 2014.
[31] Foltýnek, T., Meuschke, N., & Gipp, B. "Academic plagiarism detection: a systematic literature review", ACM Computing Surveys (CSUR), 52(6), 112. 2019.
[32] Pimas, O., Kröll, M., & Kern, R. "Know-Center at PAN 2015 author identification", Working Notes Papers of the CLEF. 2015.
[33] Ruseti, S., & Rebedea, T. "Authorship Identification Using a Reduced Set of Linguistic Features" Notebook for PAN at CLEF 2012.
[34] Satyam, A., Dawn, A. K., & Saha, S. K. "A Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis", Notebook for PAN at CLEF. 2014.
[35] Seidman, S. "Authorship verification using the impostors method", In CLEF 2013 Evaluation Labs and Workshop-Online Working Notes. 2013.
[36] Solorio, T., Pillay, S., & Montes-y-Gómez, M. "Authorship Identification with Modality Specific Meta Features", PAN, 1, 11. 2011.
[37] Vilariño, Darnes, et al. "Baseline Approaches for the Authorship Identification Task".