SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 2
ISSN 2054-7390

Authorship Identification using Generalized Features and Analysis of Computational Method

[1]Smita Nirkhi, [2]R.V.Dharaskar and [3]V.M.Thakare
[1]G.H.Raisoni College of Engineering, Nagpur University, Nagpur, India;
[2]Disha Technical Campus, Raipur, India;
[3]Department of Computer Science, University Campus, Amravati, India
smita811@gmail.com; rvdharaskar@yahoo.com; vilthakare@yahoo.co.in

ABSTRACT

Authorship Identification is being used for forensics analysis and humanities to identify the author of anonymous text used for communication. Authorship Identification can be achieved by selecting the textual features or writing style. Textual features are the important elements for Authorship Identification .It is therefore important to analyze them and identify the most promising features. This paper tries to identify and analyze promising generalized features and computational methods for authorship Identification. The performed experiments in the authorship identification task shows that, the support vector machine classifier used as computational method can achieve better results with identified generalized feature set.

Keywords: Author identification, support vector machine, feature extraction, classification

1 Introduction

Internet has provided us a platform and convenient way to share information across time and place. At the same time it is also used for criminal activities like Cyberattacks, Distribution of illegal materials in cyberspace, Computer-mediated illegal communications within big crime groups or terrorists. Cybercrime has become one of the major securities Issues for the law enforcement community. The anonymity of cyberspace makes identity tracing a significant problem which hinders investigations. Anonymity means senders will attempt to hide their true identities to void detection. Cybercriminals also Forged sender's address and Use multiple usernames to distribute online messages via different anonymous channels. Cybercrimes due to anonymity includes 1.Identity theft and masquerade 2.Phishing and spamming 3.Child pornography 4.Drug trafficking 5.Terrorism 6.Infrastructure crimes: Denial of service attacks. The possible solution for above mentioned problem is identifying the writing style of these messages. Cyber-criminal may have "word print" hidden in his online messages.

This study proposes the use of authorship analysis approach to solve the problem of identity tracing in cybercrime investigation. Problem statement is to verify whether suspect S is or is not the author of a given malicious e-mail or online message µ.with assumption that investigator has access to previously written e-mails of suspect S and have access to e-mails {E1,…,En}, collected from sample population U= {u1,…,un}.The task is to extract stylometric features and develop two models that is Training model & testing model. After that classify e-mail µ using the two models. Feature selection and computational methods are two critical research issues that influence the performance of authorship analysis. Selected

DOI: 10.14738/tmlai.32.1064
Publication Date: 28th April, 2015
URL: http://dx.doi.org/10.14738/tmlai.32.1064

features should be effective discriminators. Computational Methods provides approach to discriminating texts by authors based on the selected features. Next section of paper describes the existing work in the authorship identification field by analyzing the various features used in various research papers along with their accuracy followed by another section experimental setup which is describes the methodology used for performing experimentation. Last section shows the experimental results in terms of accuracy.

2 Related Work

Authorship analysis is categorized into three major categories [11]

1. Authorship identification (authorship attribution) which determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author
2. Authorship characterization:-It summarizes the characteristics of an author and generates the author profile based on his/her writings along with Gender, educational, cultural background, and writing style
3. Similarity detection:-It Compare multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author for e.g. Plagiarism detection.

In previous work Writing-style Features applied for Authorship Identification are Lexical features, syntactic features, Structural features & Content-specific features. Lexical features (F1) based on words and character analysis. Syntactic features (F2) perform function words, punctuation usage; POS. Structural features (F3) make use of signature, personal article-organizing style. Content-specific features (F4) analyze consistently used and content-related key words [11].Various researchers used these features for experimentation and accuracy is calculated. The table-1 shows the features and technique used for Authorship Identification along with accuracy. Chaski (2005)[8] has achieved 95.70% accuracy by using feature set (F1,F2,F3,F4) and 10 authors were used for experimentation. Similarly Iqbal (2008)[9] used Frequent Pattern Mining Algorithm to extract writing style of author. Hadjidj (2009)[7] used F1,F2,F3,F4 with accuracy 90%. Iqbal (2010)[10] used K-means with accuracy 90%. Zheng used F1, F2, F3, F4 and achieved 97.69% accuracy.

Table 1: Features used in various Research Paper

Research Paper	Number Of Authors	Features /Technique Used	Accuracy	Number Of Authors
Chaski(2005)[8]	10	F1,F2,F3,F4	95.70%	10
Iqbal(2008)[9]	10	Frequent Pattern Mining	77%	10
Hadjidj(2009)[7]	3	F1,F2,F3,F4	90%	3
Iqbal(2010)[10]	3	K-means	90%	3
Zheng	10	F1,F2,F3,F4	97.69%	10

Performance for Authorship Identification can be measure in terms of Accuracy and number of Authors used for analysis.Table-2 shows previous work done in terms of number of authors used for experimentation.

Table 2: Experimental setups from previous research.

Research Paper	Total Number of persons(P)	Total Number Of messages	Average message Length(Word)	Average Message per person
Corney et al	4	253	92	64
De Vel	3	156	259	52
Zheng et al	20	960	169	48
Stamatotes	10	300	1122	30
Tsuboi	3	4961	112	1653

From above literature survey it is observed that despite significant progress achieved on the identification of an author within a small group of individuals, it is still challenging to identify an author when the number of candidates increases .secondly, it is difficult to identify if the sample text is short as in the case of e-mails or online messages. The following experimentation shows that for short text and more number of authors, proposed methods gives more accuracy using new feature set.

3 Experimental Setup

Figure 1 shows computational methods that can be used for experimentation are divided into two categories. Those are

1. Statistical Approach uses cluster analysis and Multidimensional Scaling.
2. Machine Learning Approach uses SVM, Naïve Bays

In this paper, For Experimentation purpose method used is Support Vector Machine and Features used are most frequent words means the words with highest frequency are considered in analysis and n-gram approach.

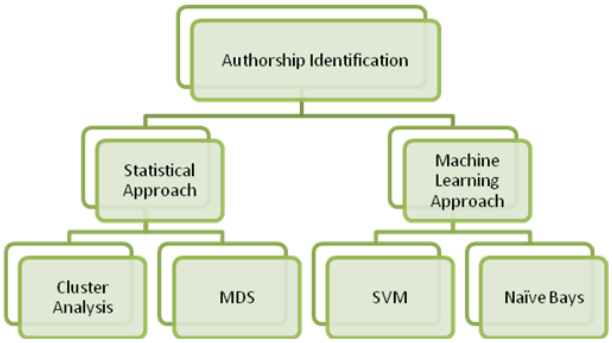


Figure 1: Computational Methods

3.1 Corpus used

3.1.1 C50 corpus

The C50 dataset was downloaded from the UCI Machine Learning Repository. It consists of one training and one test set, these sets are not overlapping. Each of the datasets contains 2500 documents (50 authors with 50 documents each) in text format. All of the documents are written in English and belongs to the same subtopic which will minimize the possibility of being able to classify documents depending on topics instead of the unique features which represent each author.

3.1.2 Enron corpus

Enron corpus was made public during the legal investigation concerning the Enron Corporation. The current version contains 619,446 messages belonging to 158 users

This dataset was collected and prepared by the CALO This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.

4 Experimental Results

Table3 shows experimental results on C50 dataset with number of authors =5, 7,15,25,50 using SVM classifier and n-gram for word=1

Table 3: Experimentation on c50 Dataset

Data set Used	Total Number Of Authors	Training set	Testing set	Accuracy
C50 Dataset	5	173	173	82.5%
	7	173	2	100%
	15	373	3	86.5%
	25	625	2	100%
	50	625	10	88%

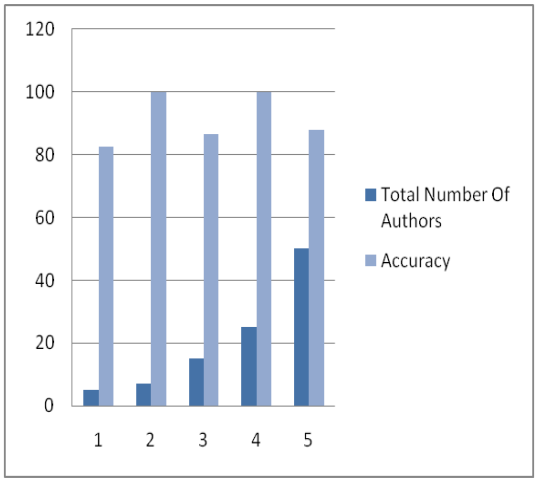
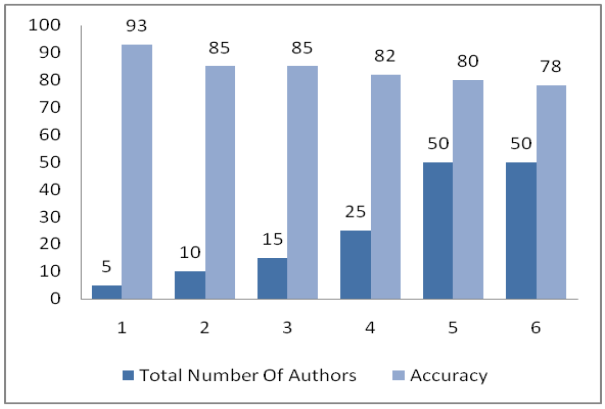


Table 4 shows experimental results on Enron dataset with number of authors =5, 10,15,25,50 using SVM classifier and n-gram for word=1

Table 4: Experimentation on Enron Dataset

Data set Used	Total Number Of Authors	Training set	Testing set	Accuracy
Enron Dataset	5	62	10	93.3%
	10	125	10	85%
	15	188	10	85%
	25	310	13	82%
	50	650	2	80%



5 Conclusion

The proposed approach is able to identify the authors of online messages. Character and word Uni-gram features showed particular discriminating capabilities for authorship identification. SVM gives more accuracy with word uni-gram .Different parameter settings of authorship identification had an impact on performance. The above experimentation shows that for short text and more number of authors, proposed methods gives more accuracy using n-gram approach for feature set.

REFERENCES

[1] Abbasi, A., & Chen, H. (2005). Analysis to Extremist- Messages, (October), 67–75.

[2] B. Loader, D.Thomas (Eds), Cybercrime: Law enforcement, security and surveillance in the information age. Routledge; 2000.

[3] A. Abbasi, H. Chen. "Writeprint: A stylometric approach to identity level identification and similarity detection in cyberspace". ACM Transaction on Information System, 26(2):1-29, 2008

[4] R. Zheng, J. Li, H. Chen, Z. Huang. "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), pp.378-393, 2006.

[5] S. Nizamani S, N. Memon N, U. K. Wiil, P. Karampelas, "CCM: A Text Classification Model by Clustering", International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Kaohsiung, Taiwan, pp.461-467, 2011.

[6] UCI Machine Learning Repositiory, Reuter 50 50 Dataset. https://archive.ics.uci.edu/ml/datasets/Reuter_50_50.

[7] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem. Towards an integrated e-mail forensic analysis framework. Digital Investigation, 5(3-4):124 – 137, 2009

[8] C. E. Chaski. Who's at the keyboard: Authorship attribution in digital evidence Investigations International Journal of Digital Evidence, 4(1), Spring 2005.

[9] F. Iqbal, R. Hadjidj, B. C. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Investigation, 5, Supplement (0):S42 – S51, 2008. The Proceedings of the Eighth Annual DFRWS Conference

[10] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation, 7(1-2):56 – 64, 2010.

[11] S.M.Nirkhi, R. V. Dharaskar, V.M.Thakre, "Analysis of online messages for identity tracing in cybercrime investigation", 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), pp. 300 - 305, 2012