ISSN:0975-9646

S.A.Aher et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1802-1805

AUTHOR IDENTIFICATION in MAIL

S.A.Aher [1],
[1] *Assistant Professor SVIT , Nashik*

D.S.Nikam [2],
[2]*BE. Student of IT, Pune*

J.M.Tadge [3],
[3] *B.E. Student of IT, Pune*

Bharti.Avhad [4],
[4]*BE.Student of IT, Pune*

V.D.Mahajan[5],
[5] *B.E. Student of IT, Pune*

***Abstract*—the cyber world provides Associate in nursing anonymous surroundings for criminals to conduct malicious activities like spamming, causation ransom e-mails, and spreading botnet malware. Often, these activities involve t matter communication between a criminal and a victim, or between criminals themselves. Authorship Identification technique are unit accustomed establish the foremost acceptable author from Gmail The forensic analysis of online textual documents for addressing the obscurity drawback referred to as authorship analysis is that the focus of most law-breaking investigations. Authorship analysis is that the applied mathematics study of linguistic and procedure characteristics of the written documents of people.**

This paper is that the initial work that presents a unified data processing answer to deal with authorship analysis issues supported the idea of frequent pattern-based whiteprint. A study of recent techniques and automatic approaches to attributing authorship of on-line messages is conferred in paper. in depth experiments on world knowledge recommend that our planned answer will exactly capture the writing sorts of people. What is more, the write print is effective to spot the author of Associate in nursing anonymous text from a gaggle of suspects and to infer linguistics characteristics of the author

***Keywords*— Author Identification, feature selection, significant features, writer identification, handwritten authorship**

I.INTRODUCTION

The speedy development and proliferation of web technologies and applications have created a brand new thanks to share data across time and area. a large vary of activities have evolved over the web, starting from straightforward data exchange and resource sharing to virtual communications and e-commerce activities. specially, on-line messages square measure being extensively wont to distribute data over Web-based channels like e-mail, Web sites, web newsgroups, and web chat rooms. Sadly, on-line messages can also be ill-used for the distribution of uninvited or inappropriate data like junk (commonly stated spamming) and o endive/threatening messages. Moreover, criminals are mistreatment on-line messages to distribute outlawed materials, together with pirated software package, kiddie porn materials, taken properties, and so on. Additionally, criminal or terrorist organizations conjointly use on-line messages united of their major communication channels. These activities have spawned the conception of crime. Crime was First State need by Thomas and Loader as outlawed computer-mediated activities which may be conducted through world electronic networks. a standard characteristic of on-line messages is namelessness. Individuals sometimes don't got to offer their real identity data like name, age, gender, and address. In several misuse or crime cases of on-line messages, the sender can commit to hide his/her true identity to avoid detection. for instance, the senders address may be solid or routed through AN anonymous server, or the sender will use multiple usernames to distribute on-line messages via completely different anonymous channels. Therefore, the namelessness of on-line messages imposes distinctive challenges to identity tracing in Internet. As results of the sheer growth of cyber users and activities, econsumer machine-driven ways for identity tracing are getting imperative.

Author Identification (AI) is one branch of pattern recognition for rhetorical application mistreatment dynamic biometric AI focuses on distinctive the writers mistreatment their individual variety of writing, as a result of the variations of distinctive important feature However, AI doesn't regarding with the which means of the word written the most issue in AI is to amass the options that mirror the author of handwriting Extracted options could embrace several garbage options Such options aren't solely useless in classification, however typically degrade the performance of a classifier so, distinguishing the many options is extremely vital role Feature choice is one in every of important analysis space for many years past The internet has become the foremost helpful platform for human action and sharing ideas Anyone will simply access the net and build comments, publish thoughts, ideas or categorical opinions Generally, most web users have an interest during a specific topic and frequently voice an equivalent opinion once moving from one page to a different the most objective of our analysis study is to propose AN approach which will establish a selected profile supported its writings on the net To observe a profile, we've grotto extract texts from the net, analyse the literary genre and therefore the vocabulary and terms used .As a neighbourhood of labour, we tend to develop a tool permitting US to observe the probable author of an internet message This methodology is use to spot the author in mail

II. REVIEW OF LITERATURE

Gray, Sallis, and MacDonell (1997) know four principle aspects of authorship analysis which will be applied to software system forensics. Supported some definitions from Grayed al., we have a tendency to categorised authorship analysis studies into 3 major fields. Authorship identification will be developed as follows: Given a collection of writings of variety of authors, assign a brand new piece of writing to 1 of them.drawback|the matter} will be thought-about as a applied math hypothesis check or a classification problem. The essence of this classification is characteristic a collection of options that stay comparatively constant for n outsized range of writings created by a similar person.

In [4] A. Anderson, M. Corney and G. Mohay have investigated the educational of authorship classes for the case of each mass and multi-topic e-mail documents. Used AN extended set of preponderantly content free e-mail document options like structural characteristics and linguistic patterns. The classier used was the Support Vector Machine learning algorithmic rule. Experiments on variety of e-mail documents generated by completely different authors on a collection of topics gave encouraging results for each mass and multi-topic author categorization. However, one author class made worse categorization performance results, in all probability because of the reduced range of documents for that author. They conjointly discovered no improvement in classification performance once as well as word collocation and even a discount in performance once the closed-class word spatial property was enhanced..

In [5] Antonio Fidel Castro Ruz the classifiers each exhibit glorious accuracy in cross validation, and do fairly well within the general use case wherever the check stream comes from a unique Twitter account by a similar author. Expected less accuracy from Twitter than from the web log or email information utilized impervious work as a result of the scale of the tweets are considerably smaller as compared. With the tiny range of tagged examples in our information set, we have a tendency to cannot build a judgment concerning whether or not performance was higher or worse on Twitter vs. blog data, however they will conclude that stylo metric analysis will, in fact, perform well on tweets.

In [2] A.K. Muda addresses that Enhanced Writer Identification Framework (EWIF) this framework is the introduction of an enhanced framework specifically for WI domain, termed as EWIF, which consists of feature extraction, feature discretization, and classification. The purpose of the feature discretization in the EWIF is to provide standard representation of individual features, which allows small variance between features for intra-class (same author) and large variance for inter-class (different authors).Although feature discretization provides better representation of individual features, this mechanism only partially reflect the key concern in WI domain, which is acquiring the features reflecting the author of handwriting, also known as unique significant features.

The main drawback of EWIF is that the mechanism to acquire the unique significant features is not present and is not defined as the part of the framework; instead the whole features are used for the identification phase. The acquisition of the unique significant features is not addressed in the EWIF. A brief framework design that summarized the phases
Incorporated in EWIF is illustrated in Fig. 1.

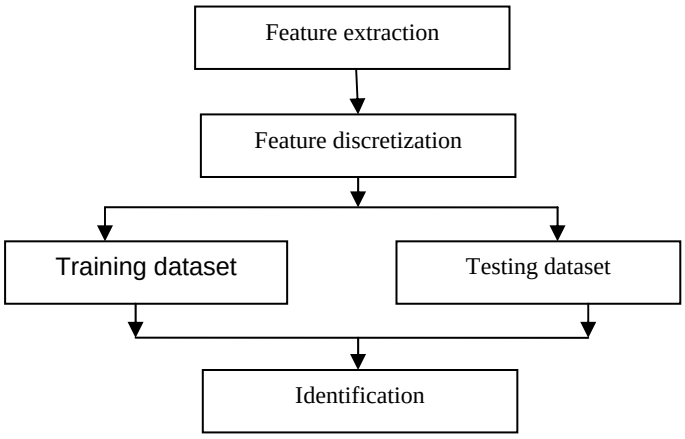

Fig 1. Framework design of EWIF.

In [6] S.F. Pratama, A.K. Muda, Y.-H. Choo, and N.A. Muda addresses the Cheap Computational Cost Class-Specific Swarm As discussed earlier, the main issue in AI is how to acquire the individual features from various handwritings [5],.

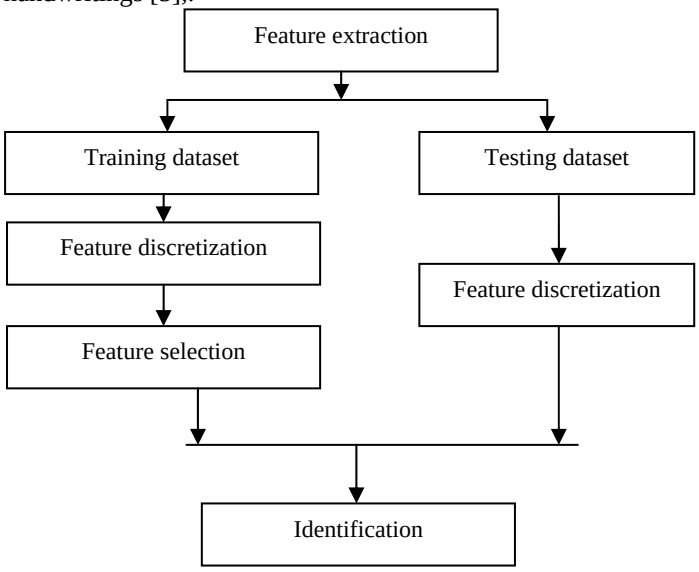

Fig 2. Framework design of C4S4

Which directly unique to those individual. Therefore, class-specific feature selection must be incorporated in order to capture these unique individual significant features. C4S4, an improvement to EWIF, caters with this class-specific feature selection issue.

The distinction between C4S4 to EWIF is that the feature discretization is conducted when the dataset has been split into coaching and testing dataset. This method is

closely representing the real-life applications, wherever the testing dataset isn't obtainable to the system beforehand and therefore mustn't been closed within the coaching method. though this method aroused another drawback completely different of various} set of information as a result of different cut-off points and intervals is used, it's solved in C4S4 by storing the discretization rules for every category, that are used throughout the classification section, wherever the testing information are discredited . C4S4 transforms the coaching information into c-binary issues by exploitation category finalization stage. After that, c-binary issues rebalanced exploitation category equalization. Feature choice stage is conducted next, that produces distinctive individual vital options. These c-feature subsets are used anon in identification stage .The framework style of C4S4 is given in Fig. 2.

III. Proposed Method

A. Framework

To overcome the limitation of existing system we have a tendency to address authorship drawback and for that we have a tendency to use a unified data processing approach that models the write print of an individual. The construct of write print, AN analogy of fingerprint in physical rhetorical analysis, is to capture the literary genre of an individual from his/her written communication. Authorship studies recommend that individual person typically leave traces of their temperament in their written work. as an example the choice of words, the composition of sentences and paragraphs, and therefore the relative preference of 1 language whole over one. Another will facilitate in characteristic one individual from another. We have a tendency to propose the event of whiteprints technique that is unsupervised technique which will be used for identification and equally detection. Write prints could be a karhunenloeve transforms-based technique that uses the window and pattern disruption.

B. Authorship Analysis

Authorship analysis could be a method of examining the characteristics of a chunk of writing to draw conclusions on its authorship. Its roots are from a linguistic analysis space known as stylometry that refers to applied mathematics analysis of literary vogue. As additional refined techniques, like machine learning techniques, are applied to the present domain, this field of analysis has been typically recognized as authorship analysis We knew four principle aspects of authorship analysis which will be applied to code forensics. Supported some definitions from we have a tendency to classified authorship analysis studies into 3 major fields. Authorship identification will be developed as. Given a collection of writings of variety of authors, assign a brand new piece of writing to at least one of them. {The drawback the matter} will be thought of as applied mathematics hypothesis take a look at or a classification problem. The essence of this classification is characteristic a collection of options that stay comparatively constant for an oversized range of writings created by constant person.

C. Parameters for Authorship Identification

In essence, authorship identification could be a classification drawback. The quality level of this drawback will be determined by many parameters. For instance, {the range the amount the quantity} of authors and therefore the number of obtainable sample documents within the coaching set might have an effect on the prediction accuracy. Most previous experimental studies of authorship identification worked on a comparatively small-scale classification drawback (i.e., 2 or 3authors). Coaching size (i.e., the full range of writings) varied wide in numerous applications. In previous literature, we have a tendency to conjointly detected small identification performance with over four authors. Though these parameters are thought of important to the quality of the matter and thus the prediction accuracy, there are not any studies examining their impact on the authorship-identification performance in an exceedingly systematic means

IV.Mathematical Model

Set theory:-

A set is de ned as a collection of distinct objects of same type on class of objects.The ob-ject of a set are called elements or members of the set. Object can be number, alphabet, names etc.

E.g.:-A=f1; 2; 3; 4; 5g

Set theory applied to the project:-

The set available are:-

Universal set (FS):-fM; S; G; Mp; Lg

Set of Messages (M):=fm1; m2; :::::::mng

Set of Suspects(S):=fs1; s2; :::::::sng

Set of Groups (G):=fG1; G2; :::Gkg

Set of Matching patterns (Mp):=fmp1; mp2; mpng

Set of Frequent stylometric patterns (L):=fx1y1z1; x2y2z2; xnynzng

Functions:-

Function shows the relationship between the elements of the set.

E.g.:-F(x)=y here y is a function of x.

There are following functions available:

(1)Grouping (F1)

(2)Extract (F2)

(3)Filter (F3)

(4)Identify Author (F4)

V.Result And Discussion

We report our results presenting the per-author-category macro-averaged F1 statistic for the Support Vector Machines (SVM) classier. The classification results are _rest presented for the case of aggregated topic categories, followed by the results for multi-topic classification. This will help us to identify the author identification in mail. We can detect the real sender of malicious mail any frauds that occur using the E-mails, websites, chat rooms or groups of the victims, to do transactions which are not legal and any scheme related to illegal it is used for obtaining and analysing digital information for use as evidence in civil, criminal or administrative cases. It contains secure

collection of computer data, the identification of suspect data, to determine details such as origin and content, the presentation of computer based information.

VI. CONCLUSIONS

We have investigated the training of authorship classes for the case of each collective and multi-topic e-mail documents. We tend to used associate degree extended set of preponderantly content free e-mail document options like structural characteristics and linguistic patterns. The classier used was the Support Vector Machine learning algorithmic rule. Experiments on variety of e-mail documents generated by totally different authors on a collection of topics gave encouraging results for each collective and multi-topic author categorization. However, one author class created worse categorization performance results, most likely owing to the reduced variety of documents for that author. we tend to additionally determined no improvement in classification performance once together with word collocation and even a discount in performance once the word Dimensionality was increased.

REFERENCES

[1] Satrya Fajri Pratama, Azah Kamilah Muda, Ajith Abraham, and Noor Azilah Muda An Alternative to SOCIFS Writer Identification Framework for Handwritten 2013 IEEE International Conference on Systems, Man, and Cybernetics .

[2] A.K. Muda, Authorship Invarianceness for Writer Identification Using Invariant Discretization and Modified Immune Classifier. Johor: Universiti Teknologi Malaysia, 2009.

[3] S.N. Srihari, C. Huang, H. Srinivasan, and V.A. Shah, "Biometric and Forensic Aspects of Digital Document Processing," in Digital Document Processing, B.B. Chaudhuri, Ed. Heidelberg: Springer

[4] A. Anderson, M. Corney and G. Mohay Mining Email Content for Author Identification Forensics

[5] Antonio Castro Author Identification on Twitter Hardesty, Third IEEE International Conference on Data Mining, 2003, pp. 705-708

[6] S.F. Pratama, A.K. Muda, Y.-H. Choo, and N.A. Muda, "SOCIFS Feature Selection Framework for Handwritten Authorship," Hybrid Intelligent Systems, vol. X, 2013, pp 83-91..

[7] A. Schlapbach, V. Kilchherr, and H. Bunke, "Improving Writer Identification by Means of Feature Selection and Extraction," Proc.Intl. Conf. Document Analysis and Recognition, 2005, pp. 131-135

[8] S.F. Pratama, A.K. Muda, Y.-H. Choo, and N.A. Muda, "Computationally Inexpensive Sequential Forward Floating Selection for Acquiring Significant Features for Authorship Invarianceness in Writer Identification," New Computer Architectures and Their Applications, vol. I, Oct. 2011, pp. 581-598.

[9] S.F. Pratama, A.K. Muda, Y.-H. Choo, and N.A. Muda, "PSO and Computationally Inexpensive Sequential Forward Floating Selection in Acquiring Significant Features for Handwritten Authorship", 11th Intl. Conf. Hybrid Intelligent Systems, 2011, pp. 358-363

[10] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. XXIII, 2007, pp. 2507-2517

[11] B.B. Pineda-Bautista, J.A. Carrasco-Ochoa, and J.F. Martinez-Trinidad, "General framework for class-specific feature selection", Expert Systems with Applications, vol. XXXVIII, 2011, pp. 10018-10024..

[12] F. Tan, Improving Feature Selection Techniques for Machine Learning, Georgia State University, 2007.

[13] K.-P. Chung, C.C. Fung, and K.W. Wong, "A Feature Selection Framework for Small Sampling Data in Content-based Image Retrieval System", ICICS, 2005

[14] S.N. Srihari, S.-H. Cha, and S. Lee, "Establishing Handwriting Individuality Using Pattern Recognition Techniques," Proc. 6th Intl. Conf. Document Analysis and Recognition, 2001, pp. 1195-1204..

[15] Z. Bin and S.N. Srihari, "Analysis of Handwriting Individuality Using Word Features," Proc. 7th Intl. Conf. Document Analysis and Recognition, 2003