A Survey of Authorship Identification in Digital Forensics based on Machine Learning Algorithms

[1] Malla Pavani, [2] D. Lalitha Bhaskari
[1]M.tech Student, Department of Information Technology,
[2]Professor, Department of Computer Science and Systems Engineering,
Andhra University College of Engineering(A),Andhra University, Visakhapatnam, India.

ABSTRACT

Authorship Identification is subfield of authorship analysis deals with finding the plausible author of anonymous messages. The Authorship identification problem of online messages is challenging task because cyber predators make use of obscurity of Cyberspace and conceal the identity. Cybercriminals make misuse of online communication for sending blackmail or a spam email and then attempt to hide their true identities to void detection. Authorship Identification of online messages is the contemporary research issue for identity tracing in cyber forensics. This is highly interdisciplinary area as it takes advantage of machine learning, information retrieval, and natural language processing. By performing the forensic analysis of online messages, empirical evidence can be collected. These evidences can be used to prosecute the cybercriminal in a court and punish the guilty. This way cybercrimes can be minimized up to certain extent by detecting the true identities. Therefore it is required to build up innovative tools & techniques to appropriately analyze large volumes of suspicious online messages. This paper compares the Performance of various classifiers in terms of accuracy for authorship identification task of online messages. Support Vector Machines, KNN, and Naïve Bayes classifiers are used for performing experimentation .This paper also investigate the appropriate classifier for solving authorship of anonymous online messages in the context of cyber forensics.

KEYWORDS*:* Authorship Identification, Cybercrime, Cyber Forensics, Support Vector Machine, K-NN, Naïve Bayes

I.INTRODUCTION

Internet provides us the convenient and efficient platform for sharing and exchanging information across the world. It has connected the world through a collective area called cyberspace. This connectivity gives many opportunities and advantages to internet users at large. However, this connectivity gives opportunity and possibility to different criminal elements for committing crimes. It is a very crucial task to identify who is behind the cyber-crime. To punish the guilty, tremendous forensics and investigation capabilities are required. Cyber-crime and cyber-attacks are increasing all over the world. In India during last 17 years conviction rate is 0.5% and cybercrime growth rate has increased to 107%.Every year there is rise in cybercrime cases.

The main cause behind these criminal activities is anonymous and decentralized nature of internet. Besides, there are many methods to hide the identity and pretending like defender . Many criminals hide themselves among legitimate internet user to commit the crime and launch cyber-attack. They compromise the computer and make it a part of botnet to perform all illegal activities. Due to available techniques of hiding the true identity it has become quite difficult to prove that a particular crime has been committed by him or her.

In this context, authorship attribution also called as Authorship identification is an important technique to detect the culprit. Data mining techniques along with the profile of accused to attribute the author of written text is the only emerging area of cybercrime investigation. In India, there is lack of regulation and guidelines for effective investigation of cybercrimes. Therefore techno legal skill development is necessary. Authorship Attribution of anonymous online messages is an upcoming research area in recent years. Earlier it has been studied in many fields like Literary and poetic work, social psychology.

Cybercrime is also known as computer crime, the use of a computer to further illegal ends, such as committing fraud, trafficking in child pornography and intellectual property, stealing identities, or violating privacy. Cybercrime, especially through the Internet, has grown in importance as the computer has become central to commerce, entertainment, and government. Senders can hide their identities by forging sender's address; Routed through an anonymous server and by using multiple usernames to distribute online messages via different anonymous channel. Author Identification study is useful to identify the most plausible authors and to find evidences to support the conclusion. Authorship analysis problem is categorized as

1) Authorship identification (authorship attribution): It 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 like Gender, educational, cultural background, and writing style.

3) Similarity detection: It compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author like Plagiarism detection. To extract unique writing style from the number of online messages various features need to be considered are Lexical features, content-free features, Syntactic features ,Structure features ,Content-specific features.

II. EXISTING METHODOLOGY

This review examines the n-gram features and machine learning techniques that are currently used in the authorship identification research area. This analysis leads to find out the current existing features and techniques that are being usedintheauthorshipidentificationresearchfield.Howcantheexistingfeaturesandtechniquesbecomparedinterms of their qualitative (accuracy level) and quantitative (number of the different person they can be distinguished and the execution time of the solution in performing that task) attributes. Over the last century and more, a great variety of Statistical & machine learning methods have been applied to authorship attribution problems of various types .

It can be divided into two classes of approach:
1. Statistical approach:
a. Unitary invariant approach
b. Multivariate analysis approach

2. Machine Learning approach: In Unitary invariant approach a single numeric function of a text is sought to discriminate between authors. In Multivariate analysis approach, statistical multivariate discriminant analysis is applied to word frequencies and related numerical features. A statistical analysis method includes cluster analysis, Multidimensional Scaling (MDS), Principal Component Analysis (PCA), consensus Tree. The machine learning approach, in which modern machine learning methods are applied to sets of training documents to construct classifiers that can be applied to new anonymous documents. The various classifiers are Delta, SVM, Naïve Bayes, and K-NN. In recent years, the research in the field of Authorship Attribution is going on very short texts and in many languages.

The challenges while Working with short texts requires robust and reliable representation of such texts as well as a Machine Learning (ML) algorithm that is able to be handled with limited data. In most studies, texts of long length are used for training phase, while studies with short text are relatively rare. If text samples are long enough it is easy to represent text features sufficiently. Reducing the length of the training samples has a direct impact on performance. This paper uses short texts between 290 and 800words pretext. This allows us to probe the scalability of the proposed approach with limited training data and very short text documents.

III. PROBLEMDESCRIPTION

There are three types of authorship analysis scenarios :

Scenario1: There are many suspects and we have to attribute the text to one of them.

Scenario2: There is one suspect and we must determine if the suspect is the author of the text or not.

Scenario3: There are no suspects. The task is to provide as much psychological or demographic information about the author as possible.

IV. METHODS FOR AUTHORSHIP IDENTIFICATION

There are various authorship attribution methods, and according to Stamatas's[4],all the methods can be classified into two groups: profile-based approach and instance-based approach.

Profile Based Approach As shown in Figure4,the profile-based approach, which is a process of concatenating all the training texts of one author and generating an author profile. The features of each author are extracted from the concatenated text. Extracted features are used in the attribution model to determine the most likely author of the dispute text. However, a profile-based approach is criticized for losing much information because of the generating profile-based feature process which is required to remove all the dissimilar contents from the same author.

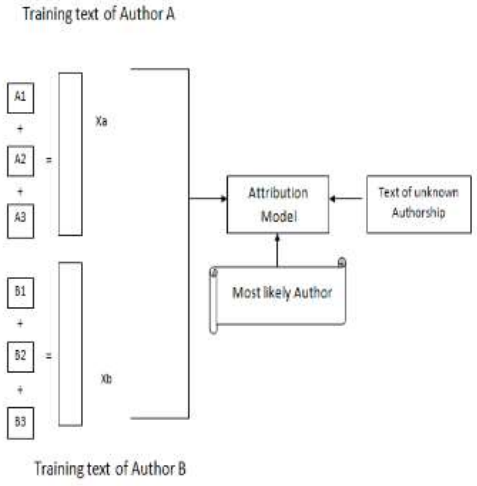


Fig. 4 Profile Based Approach

Instance Based Approach On the contrary, instance-based approach, which is used in most of the contemporary authorship attribution research, can keep most of the information from the given texts. Instance based approach, as shown in figure 5, includes every instance of training text to conclude the author of unseen text. This is discriminative approach where xa1 indicates the vector for first training sample of author A. These vectors are given as input to train the classifier and tested against the unknown authorship to find out most likely author.

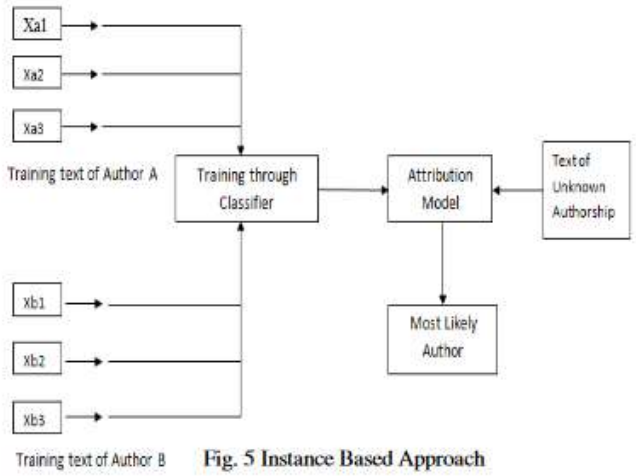


Fig. 5 Instance Based Approach

V. METHODOLOGY FOR SOLVING THE PROBLEM

Two critical research issues which influence the performance of authorship analysis is finding out the effective discriminators and approach to discriminating texts by authors based on the selected features[6]. Figure 6 shows the methodology for solving the problem. The corpus is divided into training and testing samples. Instance based approach is used. Training samples are used for Creation of Feature vector and given as input to the classifier.

Features Used

In the proposed approach, features used are Lexical Features, Character Features, unique word-print, co-occurrence of words

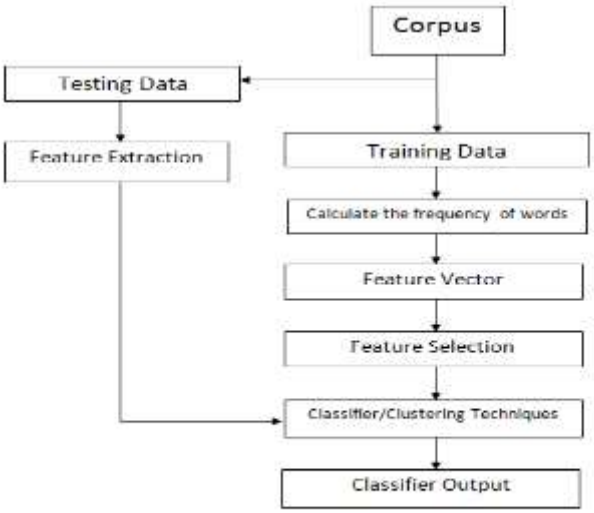


Fig. 6 Proposed Approach

1 Lexical Features includes word frequencies, word n-gram
2. Character Features are letters, digits and character n-grams.
3. Unique word-print means the list of unique words used by particular author and not by any other author. 4. Finding text style with co-occurrence of words

Classifiers Used

Authorship identification is a single label and multiclass text classification problem .Therefore selection of classifier used for performing attribution should be done carefully. This experimentation is carried out to evaluate the performance of classifiers with the selected feature sets. Classifiers used for experimentation are Support Vector Machine, Naive Bayes, and K-NN. 3.3 Corpus Used Brennan-Greenstadt Corpus is used for experimentation purpose. It was prepared at Drexel University. Two datasets are prepared by them. One dataset contain
12 authors and another contains 45 different authors. Authors were asked to submit 5000 words of their writing for demographic survey. The participated authors were known professors and friends.

VI. MACHINE LEARNING METHODS

The study in machine-learning technique is always concentrates on the selection of features in representation of document and on the selection of learning algorithms. Methods of selection are relay on whether there are two or more candidate authors. If only two candidate authors exists, then use Support Vector Machines SVMs are better descriptions of instance-based approaches are done by the vector space models. These algorithms were studied thoroughlyintheareaoftopic-basedtext-categorizationinvestigationsin.Afewofthesealgorithmscanefficiently manage multi-dimensional, noisy, and sparse data, permitting significantly numerous ways of presenting the texts. For an instance, whenever several features are used, an Support Vector Machine model is capable to avoid over fitting problems and is viewed as one of the finest solution of present technology. Class-imbalance is problem which is effecting the vector space model. A new technique was proposed by Stamatatos et.al. to handle this type of problems with the use of instance based approaches. Training set text samples can be segmented according to the size of their class. In this way several small text samples are prepared for minority authors (authors with less no. of for training sample) while

few, but lengthy, texts can be prepared for majority authors (the authors with multiple training texts).

VII. METHODS OF AUTHORSHIP ATTRIBUTION

Random Forest

It is a supervised algorithm. It is a tree based algorithm. It creates several decision trees and combines their outputs to produce a good model. The process of combining the decision trees is known as ensemble process. Advantages and Disadvantages of Random Forest. It is robust to correlated predictors. It is used to solve both regression and

classification problems. It can be also used to solve unsupervised ML problems. It can handle thousands of input variableswithoutvariableselection.Itcanbeusedasafeatureselectiontoolusingitsvariableimportanceplot.Intakes care of missing data internally in an effective manner. The Random Forest model is difficult to interpret. It tends to return erratic predictions for observations out of range of training data. For example, the training data contains two variablexandy.Therangeofxvariableis30to70.Ifthetestdatahasx=200,randomforestwouldgiveanunreliable prediction. It can take longer than expected time to computer a large number of trees

Support Vector Machines

It is a supervised learning algorithm in which given a dataset it separates them into different classes using a hyperplane. The goal of SVM is to find this hyperplane. There could be many hyperplanes but we are determined to find an optimal hyperplane. The points closest to the hyperplane in the different classes are known as support vectors and these support vectors are used to predict the classes of new data points. A new incoming point is put on the equation of the hyperplane and then is classified as to which class it belongs on the basis of which side of hyperplane it falls on the vector space. To train our machine we feed supervised data i.e. data with results already known. It learns the behavior of fraud and genuine transactions and then it can classify new transaction as to which class it belongs.

K-Nearest Neighbor

Itisoneofthemostusedalgorithmsforbothclassificationandregressionpredictiveproblems.Itsperformancedepends onthreefactors:thedistancemetrics,thedistanceruleandthevalueofK.Distancemetricsgivesthemeasuretolocate nearest neighbors of any incoming data point. Distance rule helps us to classify the new data point into a class by comparingitsfeatureswiththatofdatapointsinitsneighborhood.AndthevalueofKdecidesthenumberofneighbors withwhomtocompare.TheimportantquestionishowdowechoosethefactorK?Inordertoobtaintheoptimal value of K, the training and validation is segregated from the initial dataset. Now a graph based on the validation error curve isplottedtoachievethevalueofK.ThisvalueofKshouldbeusedforallpredictions.Wecalculatethedominantclass in the vicinity of any new transaction and classify the transaction to belong to that dominant class.

Naive Bayes

It is based upon the Bayes Theorem of conditional probability; hence it is a probabilistic model that is used for automated detection of various events. It consists of nodes and edges, wherein the nodes represent the random variables and the edges between the nodes represent the relationships between these random variables and their probabilistic distribution. We calculate predefined minimum and maximum value of probabilities of a transaction being fraud or legal. Then for a new incoming transaction we see that whether it's probability of being legal is less than the minimum defined value for legal transaction and is greater than the maximum defined value for a fraud transaction. If true then the transaction is classified as a fraud.

Logistic Regression

To combat the anomalies of linear regression where it gave values greater than 1 and less than 0, logistic regression comes into play. Despite the name being regression, LR is used for classification problems for predicting binomial and multinomial out comes, having the goal of estimating the values of parameter's coefficients using the sigmoid function. Logistic regression is used for clustering and when a transaction is ongoing it examines the values of its attributes and tells whether the transaction should proceed or not.

VIII. RESULTS AND PERFORMANCE ANALYSIS

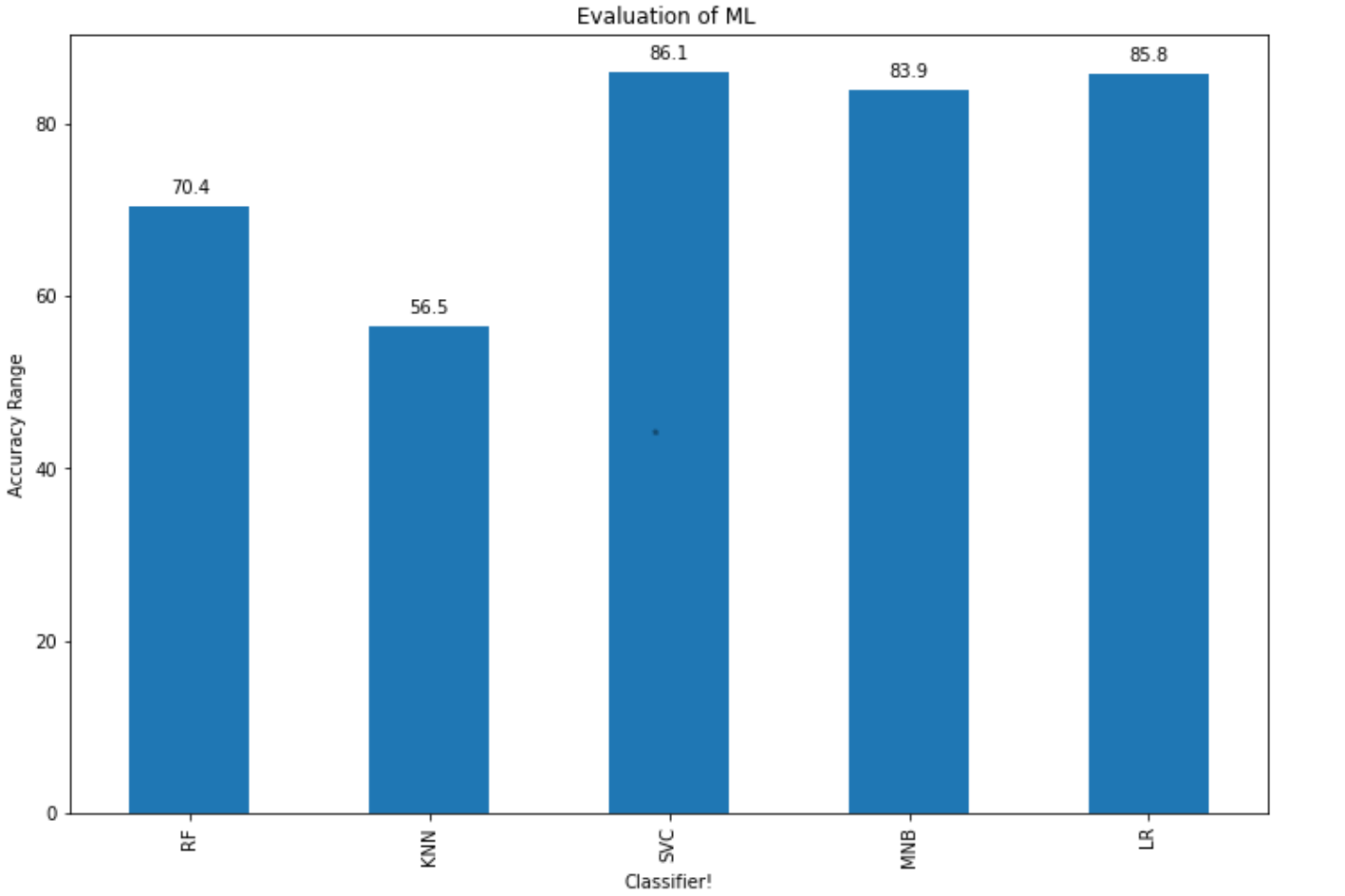
a)Results

Support Vector Machine, Naïve bayes, Logistic Regression And Random Forest classifiers are used for Classification.

For SVM, an accuracy of 86.1 is obtained, with Logistic Regression an accuracy of 85.8 and with Random Forest an accuracy of 70.4 is obtained.

b)Performance Analysis

The Performance of our implemented classifiers are shown using the histogram



CONCLUSION

A survey of various methods, classifiers and approaches were studied and presented in this paper. This paper aims at identifying the suitable model and there by analyzing the disadvantages of models in order to propose a new model. Various machine learning algorithms were studied and witnessed the applicability of these machine leaning techniques onto Authorship attribution was presented. The analysis is performed on the textual contents of messages collected online. The proposed technique used the frequency of common words from the training and testing data. Function word usage and unique word usage by each author can work as discriminator to uniquely identify the plausible author of disputed text. SVM outperforms Naïve Bays and K-NN classifiers. Different parameter settings of authorship identification had an impact on performance.

FUTURE WORK

Future work may investigate the robustness of different types of ML algorithms for tasks with many authors andsmalldatasetoftexts.Itmayalsoexpandthescopeofthestudytoinvestigateadditional(combinationsof)features.

REFERENCES

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

[2] Abbasi, A., &Chen, H. (2008). Writeprints-Level Identification and Similarity Detection in Cyberspace, 26(2). doi:10.1145/1344411.1344413

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

[4] The 9/11 Commission Report; 2002. Available online onhttp://www.gpo.gov/fdsys/pkg/GPO-911REPORT/pdf/GPO-911REPORT.pdf.

[5] Mumbai terror attacks: Telangana inaction triggered serial blasts, claims e-mail. The Economic Times; 2011. Available:http://articles.economictimes.indiatimes.com/2011-07-16/news/29781742_1_serial-blasts-ammonium-nitrate-based-terror-outfits

[6]X.Carrerars, L. S. Marquez, J. G. Salgado, "Boosting trees for anti-spam email filtering". In Proceedings of 4th International Conference on Recent Advances in Natural Language Processing (RANLP-01). TzigovChark, BG, pp.58-64, 2001

[7]H.Chen,W.Chung,Y.Qin,M.Chau,J.J.Xu,G.Wang,R.Zheng,andH.Atabakhsh.Crimedatamining:anoverviewandcase studies.InProc.Oftheannualnationalconferenceondigitalgovernmentresearch,pages1–5.DigitalGovernmentSocietyofNorth America,2003.

[8]V. D. H. Renee. Introduction to social network analysis (sna) as an investigative tool. Trends in Organized Crime, 12:101–121, 2009.ty of North America, 2003.

[9]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

[10]R.Zheng,J.Li,H.Chen,Z.Huang. "Aframeworkforauthorshipidentificationofonlinemessages:Writing-stylefeaturesand classification techniques". Journal of the American Society for Information Science and Technology, 57(3), pp.378-393,2006.

[11]T. C. Mandenhal. "The characteristics curves of composition science". Science 1887, 9(214s), pp. 237-246, 1887

[12]F.Mosteller, D.L. Wallace, "Inference and disputed authorship: the federalist". In: behavioral science: quantitative methodsedition. Massachusetts: Addison-Wesley,1964.

[13]F. Iqbal, R. Hadjidj, B. Fung, M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics." digital investigation 5 (2008): S42-S51,2008

[14]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.

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