Authorship Identification in Digital Forensics using Machine Learning Approach

Smita Nirkhi
Department of Computer Science and Engineering
G.H.Raisoni College of Engineering, Nagpur, Maharashtra, India

Dr.R.V.Dharaskar
Department of Computer Science and Engineering
MPGI, Nanded, Maharashtra, India

Dr.V.M.Thakare
Department of Computer Science and Engineering
S.G.B Amravati University, Amravati, Maharashtra, India

Abstract- Digital Forensics research field has gained the supreme importance recently due to increase in digital crimes. Many researchers are working on the various issues of digital forensics and they have developed many tools and techniques to deal with digital crimes. Digital forensics is the process of uncovering and interpreting electronic data for use in a court of law. The aim of the process is to protect any evidence in its most original form while performing a investigation by collecting, identifying and validating the digital information for the purpose of reconstructing past events. Cybercriminals make use of fake email id for attempting many cyber crimes and it's hard to identify who is the author of threatening mail or other terrorist activities. The recent research area to identify the author of such mails or online messages is Authorship identification technique, which is the one of the technique in Digital forensics helps in solving cyber crime problems. Authorship identification is a task of identifying the author of anonymous or disputed texts. An authorship identification technique helps for tracing author of anonymous text when they hide their identity through forged mail id or using proxy setting for online communication. This paper investigates the performance of existing classifiers for Authorship Identification. The accuracy of classification is depends on the selected features. Therefore this paper also investigates the various feature extraction method using n-gram.

Keywords – Digital Forensics, Authorship Identification

I. INTRODUCTION

Cybercriminals takes advantage of anonymity for performing many illegal activities like phishing, spamming, identity theft, threatening and harassment [7] [8]. In phishing, scammers able to trick account holders into disclosing their personal information like account number and password. Terrorist and criminal gangs use online messaging system as a safe channel for committing organized crimes such as armed robbers, drug trafficking and acts of terrorism. Authorship analysis research helps to find out anonymous authorship of online messages based on writing style from available samples of that author. This section reviews the research in authorship analysis. In the authorship analysis research area, there are three different sub research branches, and each one serves for a different purpose. The three sub branches are the authorship identification, authorship characterization and similarity detection.

1) Authorship identification (Authorship attribution) is used to determine the probability that a piece of writing was produced by a particular person by examining the other writings from that person.

2) Authorship characterization is a technique for determining the personal attributes of an author such as the gender, age, education level or the culture background by using existing writings from that author.

3) Similarity detection compares different pieces of writing and determines if they were produced by the same person or not. The most popular usage of this technique is the plagiarism detection.

This paper reviews authorship identification techniques using classifiers to determine the authorship of online texts and messages. Authorship Identification of online messages is a kind of classification problem where classification is based on writing style unlike to text classification where only text contents are used for classification. Authorship Identification is a task to assign text to one or more predefined classes based on the authors [3]. In recent years, Authorship Identification research witnessed a number of studies with short text. The rest of the paper is organized as follows. Existing methodology is explained in section II followed by n-gram features used for experiments in section III. Experimental results are presented in section IV. Concluding remarks are given in section V.

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 used in the authorship identification research field. How can the existing features and techniques be compared in terms 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 [4]. 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 [5]. Reducing the length of the training samples has a direct impact on performance. Traditionally, 10,000 words per author are considered to be a reliable minimum for an authorial se [6]. Some studies have shown promising results with short texts of 500 characters (Sanderson & Guenter 2006) or 500 words (Koppel et al. 2007). Siham and Halim (2012) stated that the longer is the text; the better is the identification accuracy. This paper uses short texts between 290 and 800words per text. This allows us to probe the scalability of the proposed approach with limited training data and very short text documents.

III. N- GRAM FEATURE SET

Automatic authorship identification offers a valuable technique in digital forensics for supporting crime investigation and security. It can be seen as a multi-class, single-label text categorization task. Character n-grams are a very successful approach to represent text for stylistic purposes since they are able to capture nuance in lexical, syntactical, and structural level. So far, character n-grams of fixed length have been used for authorship identification. In this paper, we investigate a variable-length n-gram approach for selecting variable-length word sequences. For experimentation we have used a subset of the new Reuters corpus [15], consisting of texts on the same topic by 50 different authors. The various character and lexical features used for experimentation are listed below in Table1.

Table 1. Character and lexical features used in this study

Feature	N-grams level	Description	Feature Type
Character Unigram	1-gram	individual characters	Character
Character Bi-gram	2-grams	two consecutive characters	Character

Character Tri-gram	3-grams	three consecutive characters	Character
Word Uni-gram	1-gram	single words	Lexical
Word Bi-gram	2-grams	2 consecutive words	Lexical
Word Tri-gram	3-grams	3 consecutive words	Lexical

IV. EXPERIMENTAL RESULTS

AA (Authorship Attribution) process starts with data pre-processing, followed by feature extraction, classification and finally author identification. In this paper, AA is considered as a classification task. Where most of Text Classification systems apply two stages approach which first extracts features with high predictive value for the classes, then it trains an ML (Machine Learning) algorithm to classify new documents by employing the selected features in the first stage. Automatic TC labels documents according to a set of pre-defined authorship classes. In the first phase, predictive features are extracted from the data, after that training and test instances are created, on the basis of these features. In the second phase, an ML model is built from training data, so as to be tested on unknown test data. The training and test instances are numerical features vectors that represent term frequency of every selected feature, followed by the author label. Also the task of AA here is conducted as multi class AA.

The performance of the classification algorithms with different selected features on the above mentioned dataset is evaluated by looking at standard evaluation metrics. Accuracy is used to indicate the number of correctly classified instances over the total number of test instances by calculating the average of accuracy, as in Eq.(1).

$$\text{Accuracy}= \frac{\text{Number of documents that are correctly classified}}{\text{Total Number of documents}} \qquad (1)$$

Machine Learning Techniques used in our experiment are SVM and KNN. Table 2 shows the performance of each machine learning technique.

Table2: The accuracy percentage by applying different features to SVM, KNN classifier.

Feature	Accuracy of SVM Classifier	Accuracy Of KNN
Character Unigram	72.9%	55.9%
Character Bi-gram	88.2%	74.2%
Character Tri-gram	84.7%	74.2%
Word Uni-gram	93.3%	70.3%
Word Bi-gram	81.7%	53.5%
Word Tri-gram	60%	40.9%
Average Accuracy	80.13%	61.5%

V.CONCLUSION AND FUTURE WORK

In this paper, an Authorship Attribution task has been experimented on a Reuter dataset which consist of 50 authors with 50 texts per author, which means total 2500 samples for training and 2500 samples for testing. Several state of the art features have been tested on this dataset. The classifier that has been implemented is K-NN and SVM. Experiments of AA have been done separately for character and Lexical feature on a Reuter dataset using an SVM classifier shows the following remarkable points:
1. The character-based features are better than the word-based features, depending on the average accuracy of all character-level features compared to the average accuracy of all word-level features of the Reuter dataset.
2. The word unigram features gave the best score for this classifier obtained up to 93.3% classification accuracy.
3. The NB classifier shows good performance in this experiment of AA. This is because the average accuracy percentage obtained the score of 71.85% (score of all used features). When compared to SVM which obtained average accuracy percentage of 62.96%.

Future work may investigate the robustness of different types of ML algorithms for tasks with many authors and small dataset of texts. It may also expand the scope of the study to investigate additional (combinations of) features.

REFERENCES

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

[2] Abbasi, A., & Chen, H. (2008). Writeprints : A Stylometric Approach to Identity-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 on http://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). Tzigov Chark, BG, pp.58-64, 2001

[7] H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, R. Zheng, and H. Atabakhsh.Crime data mining: an overview and case studies. In *Proc. Of the annual national conference on digital government research*, pages 1–5. Digital Government Society of North 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. "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.

[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 methods edition. 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.