Survey: Author Identification using Data Mining Techniques

Aishwarya Chavan[1] Yashodhara Jamdade[2] Priyanka Hundalekar[3] Poonam Kamble[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Modern Education Society's College of Engineering, Pune, Maharashtra, India

Abstract— With the rapid growth of the Internet and the expansion of its users, the Internet is becoming an ideal platform for the criminal activities which mainly includes committing fraud, stealing identities, or violating privacy, and etc. The sender can hide their true identity by creating sender's address; Route through an anonymous server and by using multiple usernames via different anonymous channel and perform the crime. This type of the message spread the incorrect information and evidence in society that give rise to social conflict. This paper presents a survey of the literature on author identification schemes and different techniques up till date. The paper outlines an overview of the author identification schemes. The different algorithms adopted for the schemes are discussed. Outlining the application and merits of the same.

Key words: Cybercrime; Author Identification; Naïve Bayes Classifier; SVM

I. INTRODUCTION

In his domain a published author generally has unique writing styles. In this field researchers believe that the writing styles of every author can varies according to their word choices, sentence structures, etc. The process of identifying the author from a group according to his writing samples (sentences, paragraphs or short articles) is authorship identification. Authorship identification is a research area which focuses on the relationship between writers and their writings. With the progression in the authorship analysis research field, different features and techniques have been developed in order to support for the research.

An approach to intelligent author categorization has been proposed using a Naive Bayes and SVM classification algorithm. The categorization of author is not only based on the body of text but also on the header of a text or article. The metadata produces an additional information that can be exploited and enhance the categorization capability. Result of system for real text data categorization is 'author names'. SVM works on features to identify author.

To identify the most feasible authors and to find evidences to support the conclusion author identification study is most beneficial. Automating authorship identification assures more accurate results and objective measures of reliability, both of which are unfavorable, for legal and security applications. In our work, two labeled datasets are adapted to train and test our models. Naive Bayes and SVM Classification algorithms are utilized at different levels to evaluate the accuracy of authorship identification. Author Identification using Naive Bayes system helps for classify text into author category.

II. LITERATURE REVIEW

[1]The author focuses on short texts retrieved from Twitter (www.twitter.com), a social networking site that limits users to 140 character messages, commonly referred to as "tweets". It also examines potential avenues of author identification in Twitter using supervised learning methods for data classification. Specifically, experiments were conducted using Support Vector Machines (SVM) with a variety of feature set options.

[2]The main objective of this thesis is the ability to detect the authorship of a research paper by using different classification algorithms and see how they perform. In this scenario, the author can use authorship identification software to check whether or not the author can be identified. If identified the author can alter the contents of the paper rendering the software unable to correctly identify the authorship and, therefore, be able to get an unbiased opinion on the work.

[3]The contributions of this paper are summarized as follows: _ the authorship identification is performed on both a news dataset (Reuters 50_50) and a story dataset at both sentence level and article level.

Gated Recurrent Unit (GRU) network and Long Short Term Memory (LSTM) network are implemented, tuned and evaluated on the performance of authorship identification.

_ Siamese network is proposed to examine the similarity of two articles. It proves to be powerful on authorship verification.

[4]The key idea of author is to determine which types of information make it possible to identify the author of a short digital text. In particular, is it possible to identify the author of a short text based on the words and grammar used? This is actually a classification task. The features of the text decide to which author (category) the text belongs.

[5]Author Identification study is useful to identify the most plausible authors and to find evidences to support the conclusion. When an author writes they use certain words unconsciously and we should able to find some underlying pattern for an author's style. The fundamental assumption of authorship attribution is that each author has habit of using specific words that make their writing unique Extraction of features from text that distinguish one author from another includes use of some statistical or machine learning techniques.

[6]The author has introduced a stylometric representation learning approach for AA. The goal is to learn an effective vector representation of writing style of different linguistic modalities in AA study. Our design inherits the flexibility of the original hand-crafted stylometric features while it enables the representation to be learned from the available data.

[7]The author has described the three implementation phases. Depending upon the frequency of users visiting each page mining is performed. By finding the session of the user we can analyze the user's behavior by the time spend on a particular page.

Some algorithms used for Author Identification are:

A. Naive Bayes Classification Algorithm

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability. To build machine learning models particularly for disease prediction and document classification. Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content. An algorithm is said to be naive when it is simple and straightforward but does not exhibit a desirable level of efficiency despite finding a correct solution or it does not find an optimal solution to an optimization problem, and better algorithms can be designed and implemented with more careful thought and clever techniques. Naive algorithms are easy to discover, often easy to prove correct and often immediately obvious to the problem solver. They are often based on simple simulation or on brute force generation of candidate solutions with little or no attempt at optimization. Despite their inefficiency naive algorithms are often the stepping stone to more efficient, perhaps even asymptotically optimal algorithms, especially when their efficiency can be improved by choosing more appropriate data structures. Naive Bayes classifiers are highly scalable requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not a Bayesian method.

1) Use of Naïve Bayes Classifier

1) If you have a moderate or large training data set.
2) If the instances have several attributes.
3) Given the classification parameter, attributes which describe the instances should be conditionally independent.

2) Naïve Bayes Features

– Intended primarily for the work with nominal attributes

In case of numeric attributes:

1) Use the probability distribution of attributes (Normal distribution is default) for probability estimation for the each attribute.
2) Discretize the attribute's values.

B. Support Vector Classification Algorithm

Support vector machine (SVM) proposed by vapnik and Cortes have been successfully applied for gender classification problems by many researchers. In its simplest form, a support vector machine is visualized in two-dimensional space with a dataset that consists of two different classes, i.e. a square and a circle. These squares and circles are separated by a hyper plane, and since the support vector machine is two-dimensional, the hyper plane is represented as a one-dimensional line. An SVM classifier is a linear classifier where the separating hyper plane is chosen to minimize the expected classification error of the unseen test patterns. SVM is a strong classifier which can identify two classes. SVM classifies the test image to the class which has the maximum distance to the closest point in the training.

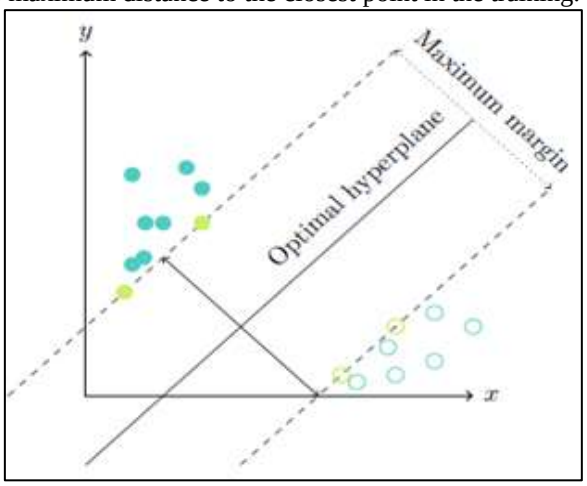

Fig. 1: Example of an SVM.

SVM training algorithm built a model that predict whether the test image fall into this class or another. SVM require a huge amount of training data to select an affective decision boundary and computational cost is very high even if we restrict ourselves to single pose (frontal) detection. The SVM is a learning algorithm for classification. It tries to find the optimal separating hyper plane such that the expected classification error for unseen patterns is minimized.

For linearly non-separable data the input is mapped to high-dimensional feature space where they can be separated by a hyper plane. This projection into high-dimensional feature space is efficiently performed by using kernels. More precisely, given a set of training samples and the corresponding decision values {-1, 1} the SVM aims to find the best separating hyper plane given by the equation WT x+b that maximizes the distance between the two classes.

III. APPLICATIONS & ADVANTAGES

A. Application

– Application may be any social media application or chatting application.
– To provide effectiveness for author identification.

B. Advantages

– Text classification using Naive Bayes system helps for author identification.
– Helps to detect cybercrime.

IV. CONCLUSION

By analyzing the literature survey we come to know that the proposed system gives the accurate result compared with other methods. As we are using large dataset which will ensures the better performance compared as earlier. Thus we build up an intelligent author identification system using a Naive Bayes and SVM classification algorithm.

By taking the advantage of the extensibility of the system, use of other classifiers that can increase the accuracy of the system in future. In addition working on other different

feature attributes which make the system more efficient and gives author identification.

REFERENCES

[1] Rachel M. Green and JohnW.Sheppard."Comparing Frequency- and Style-Based Features for Twitter Author Identification". Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference
[2] SimenSkoglund "Authorship Identification of Research Papers".Norwegian University of Science and Technology, Department of Computer and Information Science. August 2015.
[3] Chen Qian, Tianchang He, Rao Zhang "Deep Learning based Authorship Identification" Department of Electrical Engineering Stanford University, Stanford, CA 94305.
[4] Marcia Fissette and dr. F.A. Grootjen. "Author identification in short texts". 2010
[5] SmitaNirkhi and Dr.R.V.Dharaskar "Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.5, 2013.
[6] Steven H. H. Ding, Benjamin C. M. Fung, Senior Member, IEEE, FarkhundIqbal, and William K. Cheung."LearningStylometric Representations for Authorship Analysis".2168-2267 _c 2017 IEEE.
[7] G. Neelima and Dr. SireeshaRodda. "Predicting user behavior through Sessions using the Web log mining". International Conference on Advances in Human Machine Interaction (HMI - 2016).
[8] Jihoon Yang_ and Sung-Yong Park "Email Categorization Using Fast Machine Learning Algorithms" _Springer-Verlag Berlin Heidelberg 2002.
[9] James Conigliaro "Author Identification Using Naïve Bayes Classification"
[10] Sreeraj.M and Sumam Mary Idicula "A Survey on Writer Identification Schemes" International Journal of Computer Applications (0975 – 8887) Volume 26– No.2, July 2011.

