Author Identification using SVM and Naive Bayes Techniques

[#1]Tejaswi Pisal, [#2]Shraddha Shah, [#3]Aarti Nagawade, [#4]Rutuja Kakade,
[#5]Prof. Anita Gunjal

[#1234]Department of Computer Engineering,
[#5]Professor, Department of Computer Engineering,

Maharashtra Institute of Technology College of Engineering
(*Affiliated to Savitribai Phule Pune University*) Pune, India.

Abstract :With increasing use of social media, cybercrime cases are happening rapidly. Social crises may occur due to wrong messages/posts. Identities of wrong messages/post are hidden by sender. An approach to intelligent author categorization has been proposed using a Naive Bayes and SVM classification algorithm. The categorization is based on not only the body but also the header of a text or article. The metadata provide additional information that can be exploited and improve the categorization capability. In particular, categorization based only on the header information is comparable or superior to that based on all the information in a text or message. SVM works on features and we will work on new features to identify author where as Naive Bayes is used for classification.

Keywords— ***Feature Extraction, Support Vector Machine Algorithm, Nave Bayes classifier, Author Categorization, Author Identification***

I. INTRODUCTION

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 etc. With increasing use of social media usage, cybercrime occurs frequently. The sender can hide their true identity and do the crime. This type of the message spread the wrong information in society. It causes some social conflict. In some cases, it is difficult to find out the sender of that information or track the sender. The main problem is that how to identify the sender details. To overcome this problem we proposed a system which can help to identify the author. Author identification is a research area that emerged out of the increased use of internet. It is also used for determining which author wrote a chapter or passage of a book, the bible being the most famous example. Author identification research makes use of the structure of the text and the words that are used.

The sender can hide their true identity by creating senders 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.

An approach to intelligent author categorization has been proposed using a Naive Bayes and SVM classification algorithm. Generally author has unique writing styles in his domain. 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 writing styles. With the progression in the authorship analysis research field, different features and techniques have been developed in order to achieve greater accuracy in the research.

In proposed system, 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. RELATED WORK

Most of the research work has been carried out for author identification over half a century. Aautomated methods have been studied. Here survey is done on different techniques used for author identification.

Rachel M et al.[1] 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.

Simen Skoglund et al.[2] gives the system that gives 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 the author can be identified or not. 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.

Chen Qian et al.[3] has given deep learning based Authorship Identification. The contributions of this paper are summarized as: i) The authorship identification is performed on both a news dataset and a story dataset at both sentence level and article level. ii) Gated Recurrent Unit (GRU) network and Long Short Term Memory (LSTM) network are implemented, tuned and evaluated on the performance of authorship identification. iii) Siamese network is proposed to examine the similarity of two articles. It proves to be powerful on authorship verification.

Marcia Fissette et al.[4] 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.

Smita Nirkhi et al.[5] propose a Author Identification study which is 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 it should able to find some underlying pattern for an authors 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.

Steven H. H. Ding et al.[6] presented 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. It inherits the flexibility of the original hand-crafted stylometric features while it enables the representation to be learned from the available data.

[7] In this paper, they have implemented three 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.

Table 1. The summary of literature review.

Author	Feature types	Classification type	Accuracy	Goals of study
Ragel et al.(2013)	Unigrams	Cosine similarity and the euclidean distance	25%	Identification
Stamatatos 2007	Common n- gram	Svm	70%	Identification
Layton, et al. 2010	Character n-grams	Scap Algorithms	70%	Identification
Steyvers, et al.2004	Author-topics and topic word models	Svm	72%	Topic discovery
Iqbal, et al. 2010	Lexical, syntactic, and structural	Bayesian network	80.6%	Authentication
Tan, et al. 2010	13 syntactic and 4 lexical	Naive bayes	81.98%	Identification
Pavele, et al.2009	Conjunctions and adverbs	Prediction by partial matching (ppm), and svm	83-86% for ppm 82.9- 84% for svm	Identification
Monaco ,et al.2013	Lexical and syntactic	K-nn	91.5%	Authentication
Howedi et al.2014	Lexical, stuctural,syntactic and content specific character n-gram	Naive bayes and svm	96%	Identification

III. PROPOSE FRAMEWORK

The proposed system can identify the author from the given text data. Here the system can classify the text into different author category and using Naive Bayes and SVM classification algorithm. SVM may used for extract the feature such as number of tokens, number of sentences, fraction of punctuations, total length of sentences, etc and at the same time Naïve Bayes creates the bag of words as the task of data preprocessing. In proposed system, preprocessing techniques used are Removal of special symbols, Removal of stop words, and finally applying stemming. After preprocessing, Classification is done by SVM as well as Naïve Bayes algorithm. As the input text is given to both the algorithm i.e SVM and Naive Bayes algorithm, the final output may be produced by comparing accuracy created by both algorithm.

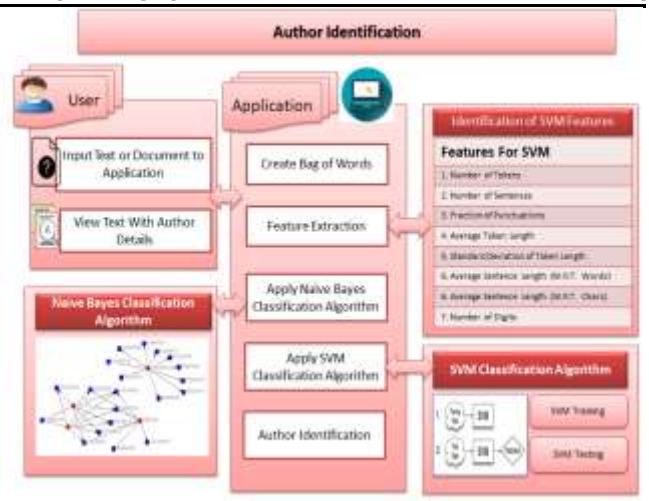

Fig 1.Process of Author Identification.

IV. METHODOLOGIES

In the proposed system, dataset of 50 author's will be using for training and testing. For each author, 50 text documents written by him/her will be used to obtain good results on a different classifier. Initially, data is processed using Preprocessing techniques. Firstly, removal the special symbols such as punctuation marks and then removal the stop words such as "a", "the", "and" etc will be done and lastly stemming will be applied to reduce the word to its root form. After preprocessing of the text, it is given to the SVM and Naive Bayes classifier for author identification.

A) Support Vector Machine (SVM): Support Vector Machines is classification technique that analyzes data. SVM solves classification problem in less time. It uses a flexible representation of the class boundaries. It can also solve a variety of problems with very less or many parameters. It can separate a set of samples possessing different classes. This technique is used for feature extraction. Extract features from author written text for calculate uniqueness for author identification. Following features are considered for identification writer.

- Number of tokens,
- Number of Sentences,
- Fraction of punctuation,
- Average token length,
- Standard deviation of token lengths,
- Average sentence length(wrt words),
- Average sentence length(wrt chars) and
- Number of digits.

Following features extracted from text in author identification are defined as:

1. Number of tokens:

A token is a sequence of characters that can be treated as a unit in the grammar of the programming languages. It is calculated as:

Number of tokens = total number of characters in the given text

2. Number of sentences

Number of sentences is calculated as total number of sentences in the given text which is terminated by a dot.

3. Fraction of punctuations

It helps the reader to understand a sentence through visual means other the simple letters of alphabets. It calculate the number of punctuations in a given text.

4. Average token length

Average token length is the ratio of total length of token to the count of token in the text.

ATL=total length/count of token

5. Standard deviation of token lengths

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

6. Average sentence length (wrt words)

Average sentence length is the ratio of total length of sentence with respect to the word to the count of sentences in the given text.

ASL=total length/count of sentence wrt words

7. Average sentence length (wrt chars)

Average sentence length is the ratio of total length of sentence with respect to the characters to the count of sentences in the given text.

ASL=total length/count of sentence wrt characters

8. Number of digits

Number of digits present in the given text

B) Naïve Bayes : The Naïve Bayes (NB) classifier is a probability-based approach. It is a powerful algorithm for predictive modeling. This probabilistic classifier has secure independent assumptions as it is based on Bayes Theorem. It is generally used in text classification and mainly with multiple classes' problems. It uses less training data as compared to other machine learning techniques. When dealing with the text, it is quite common to transform the text into data that can be easily analyzed and quantify. So Bags of model is used to represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. After transforming the text into "bag of words", we can calculate various measures to characterize the text. The common type of characteristics calculated from the Bag-of-words model is term frequency, i.e the number of times a term appears in the text.

Term Frequency-Inverse Document Frequency(TF-IDF): It is a numerical statistic that reflect how important a word is to a document in a collection . The term frequency is typically defined as the number of times a given term t appears in a document *d* (this approach is sometimes also called raw frequency. It is given as

$$TF(t)=\frac{\text{Number of times term 't' appears in a document}}{\text{Total no of terms in the document}}$$

Inverse document frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word.

$$IDF(t)= \log e \quad \frac{\text{(Total no of documents)}}{\text{(Number of documents with term 't' in it))}}$$

The Term frequency-inverse document frequency is another alternative for characterizing text documents. It can be understood as weighter term frequency, which is especially useful if stop words are not been removed from the text. The Tf-idf is assumes that the importance of word is inversely proportional to how often it is occurring in the document. It is most commonly used to rank documents.

Tf-idf=TF * IDF

V. CONCLUSION

An approach to intelligent author categorization has been proposed using a Naive Bayes and SVM classification algorithm. According to the input data the accuracy may be calculated by both the algorithm. The SVM algorithm may gives better accuracy for short text, where as Naïve Bayes works better for long text. In our proposed system, both the algorithm work in parallel manner and whose accuracy is better may given as a output. Hence the system for real text data categorized into author name.

REFERENCES

[1] Rachel M. Green and John W. Sheppard. "Comparing Frequency- and Style-Based Features for Twitter Author Identification." Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference.

[2] Simen Skoglund "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]Smita Nirkhi 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, Farkhund Iqbal, and William K. Cheung. "Learning Stylometric Representations for Authorship Analysis" 2017 IEEE.

[7]G. Neelima and Dr. Sireesha Rodda. "Predicting user behavior through Sessions using the Web log mining" International Conference on Advances in Human Ma- chine Interaction (HMI - 2016).