Cluster Computing (2019) 22:S12101–S12110
https://doi.org/10.1007/s10586-017-1563-3

An improved framework for authorship identification in online messages

L. Srinivasan[1] · C. Nalini[2]

Received: 7 September 2017 / Revised: 14 December 2017 / Accepted: 15 December 2017 / Published online: 28 December 2017

Abstract

The authorship identification will determine the likelihood of the writing produced, by an author, by means of examining the other writings. The rapid proliferation of technologies along with the applications of the internet, the misuse of online messages for the purpose of inappropriate or for illegal reasons is a major concern in society. The online message distribution and its anonymous nature will make the identity of tracing anyone of critical issue. The work has been developed using a framework for the identification of authorship of the online messages for addressing as well as tracing such problems. For this framework, identification of authorship is done by the four writing style features (the lexical, the syntactic, the structural, and the n-gram features) that are extracted and inductive learning algorithms have been used for building a feature based classification model for the identification of the authorship of the online messages. For this work, the C4.5, the fuzzy and also the Ada boost classifiers will be used for the task of authorship-identification. An experimental study on this framework with the effects of these classification techniques on online messages is evaluated.

Keywords Authorship identification · Online messages · Lexical features · Syntactic features · Structural features · Ngram features · C4.5 · Adaboost and fuzzy classifier

1 Introduction

Internet has seen a rapid development as well as have paved a new way in sharing information over time and space. There are a wide range of activities that have evolved over the internet that range from simple exchange of information and sharing of resources to the virtual communications and the activities of e-commerce. Particularly, the online messages are used extensively for distributing the information over the channels that are web-based like the websites, the newsgroups, the chat rooms and the e-mail [1].

But unfortunately, the online messages can also get misused by the distribution of inappropriate information or unsolicited information like the junk mail known as spamming and offensive messages. And the criminals also use online messages for distributing the illegal materials that include pirated software, stolen property, materials of child pornography and the like. Additionally, some criminal or terrorist organizations can also make use of online messages for communication. Such activities have also spawned the cybercrime concept. Cybercrime [2] has been defined as the activity that is made illegally and is conducted by using a global electronic network.

Another important trait of the online messages will be anonymity. People normally do not provide real identity like name, gender, age or address. Many may even misuse them or there are cases of crime and the sender makes an attempt to hide his identity. For instance, the address of the sender can also be forged for avoiding detection. So, the online message anonymity has imposed a unique challenge for identifying and tracing in cyberspace. The result of the sheer growth of the cyber users and their activities by efficient automated methods for tracing identity have now become imperative. The identification of authorship is based on the analysing of stylistic features of the online messages suggested as a possible solution.

The Authorship identification is that task of identifying authors of various anonymous texts, in accordance to the given number of examples of writing of a set of candidate

✉ L. Srinivasan
srinivasan_l@yahoo.com; srinivasanl.1982@gmail.com

C. Nalini
nalinic48@yahoo.com; nalinikec@gmail.com

[1] Dhirajlal Gandhi College of Technology, Salem, India

[2] Kongu Engineering College, Erode, India

authors that are predefined. The initial work of the authorship identification has been to attribute the authorship to the work of the Shakespearean plays from the nineteenth century. In the recent years, the large volume of such anonymous texts like online forum messages, source codes, blogs and emails are necessary for the identification of authorship [3].

This identification is to be applied to more applications that include intelligence, civil law, criminal law, computer forensics and literary works. There is also another important role in retrieval of information, extraction and answering of question. In literature, applying authorship identification is illustrated by means of identification of the work of authors or disputed authorship like that of The Federalist Papers. Recognition of writers of threatening or offensive messages has been discussed in the applications of criminal law. In addition to this, there is an example in the forensic applications of computers for judging the programmers' identity and the source codes that can destroy the data or the computers.

This task focusses mainly on two different issues. One being how to extract features of the texts that represent different styles of writing of different authors and how the appropriate methods are chosen for predicting the authors of the texts. The text representation will have features known as style markers that are objective, content independent and quantifiable for different types of texts. The features that are stylometric, used in the current features are divided into six: the character, the lexical, the syntactic, the structural, the semantic and the application-specific features.

The Authorship analysis is found to be useful in the application context in which there is an authorship attribution (AA) that is uncertain, unknown, or even otherwise obfuscated. Such type of occurrences will often arise in those disciplines like history and even criminology. Traditionally, the authorship analysis is performed through a manual analysis. However, such manual analysis is now increasingly difficult with the growing usage of electronic texts as well as social media. The problems in manual analysis will arise when large volumes are processed and traditional stylometric analysis of text content is done across several languages. By means of borrowing the techniques and perspectives from the computational linguistics there are several traditional features that are used for the purpose of evaluating the authorship that is operationalized for usage in electronic texts [4].

Particularly, there seems to be a great interest in the development of authorship visualization tools that support greater accountability to users in the online communities as well as social media. This anonymity that is given by the internet has made it very attractive as a platform for conducting different forms of crime that include trafficking of drug, piracy, terrorism and cybercrime. In addition to this, there are also other trust issues that have an online deception among individuals as well as organizations that may be mitigated using better services for authentication and visualization tools will help in deterring abuse.

The AA is that domain that aims at recognizing the author of any given text sample and one important AA challenge is that of the identification of the people that are involved in the chat conversations. This task is much more important once there is a penetration of the social media in everyday life of people having the possibility of hiding themselves behind nick names or fake profiles. Thus, far the standard stylometric features are employed for categorizing the chat content and the participants behaviour and any attempts to identify such search participants are still few and the similarity in the spoken conversation as well as chat interaction is neglected and the key difference between written information and chat data [5].

The Cybercriminals take advantage of the anonymity for the performance of several illegal activities like theft, harassment, threatening, phishing and spamming. In phishing the scammers can trick the account holders and make them disclose personal information like account number and password. The criminal gangs and terrorists use online messaging as a safe channel for the purpose of committing many organized crimes. The authorship analysis research will help in finding out the authorship of several online messages that are based on the style of writing from the available samples of the same author. In this analysis there are three sub research branches and each of that serves for many different reasons. The sub branches are the identification of authorship, characterization and detection of similarity [6].

The authorship identification (AA): this determines the likelihood by a small a piece of writing that is produced by an author by an examination of his other writings.

The authorship characterization: this summarizes the author's characteristics and generates the profile of the author that is based on his or her writings like the writing style, cultural background, education or gender.

The similarity detection: this compares several pieces of writing as well as determination of whether they are produced using a single author with no identification or plagiarism detection. For extracting the style of writing there are many online messages that are considered like the Lexical features, the content free features, the syntactic features, the structured and content specific features.

Even though the problem of AA is studied in history, in the last few decades, this is a forthcoming area of research. In the initial stages this problem started as a basic problem of identification of author of anonymous texts that are taken from Marlowe, Bacon, and Shakespeare, now is grown into forensic analysis and electronic commerce. This is an extended version of the problem of author attribution that is defined as a problem of needle-in-a-haystack [7].

An Authorship identification may be formulated as below: when a set of writings of some authors is given a new piece of writing. This problem is considered as one such statistical test of hypothesis for a problem of classification. The main essence of this is to identify the features which normally remain constant. As soon as the feature set is selected a given writing will be created and represented by the n-dimensional vector, in which n denotes the total number of features and for a set of vectors that are pre-categorized it can also apply certain analytical techniques for determining the categories of some new vectors that are created on the basis of a new writing. Therefore, feature sets and their analytical and can affect the authorship identification performance and also the issue of multiple language being important in the new direction of research.

This work proposes a new framework of identification of authorship on the online messages and for this framework there are four different types of features which have been identified in the research of authorship analysis that has been extracted from the online messages. Owing to the Internet being international the applicability of this proposed framework for various languages has been evaluated.

If an author writes some words unconsciously will find an underlying pattern for the style of the author. The basic assumption of the AA is that the author has a habit of certain words that makes a unique feature extraction that distinguished one author from that of another. The classification is now a supervised learning issue which accepts a labelled data and creates a classifier that sorts the data into various predetermined classes. This issue can be resolved through the C4.5, the adaboost and the fuzzy classifier. Here an improved identification of online messages has been proposed. Section 2 deals with literature that is relevant, Sect. 3 discuses methods used, Sect. 4 exposes the outcome and Sect. 5 concludes the work.

2 Related works

Halvani et al. [8] described K-nearest neighbor (K-NN) that is based on the authorship verification method for author identification (AI) and task of the PAN 2013 challenge. This method also follows an ensemble classification technique that is based on a combination of suitable categories of features. For each of the chosen feature category that is applied to a K-NN classifier is calculated as a style deviation score between the training documents that is calculated from a true author A and one that claims to be A. There are many benefits here like the linguistic resources like the ontologies, the language models and the thesauruses. Also, the method is extended or even modified using some parameters.

Zamani et al. [9] further introduced authorship identification as an important part in law and journalism which is a major technique in detection of plagiarism. Here for tackling the verification of authorship, proposed a probabilistic distribution for being represented as each of the documents as a feature set for increasing the interpretability of the features and their results. This also brought about a distance measure for computing the distance between both the sets. Lastly, a K-NN based approach is exploited and a dynamic method of feature selection that discriminates the writing style of the author. This also shows that this selection is needed for outstanding performance.

Qian et al. [10] has presented a novel two-view framework of co-training for identifying iteratively the authors of some unlabelled data for augmenting the training set. The main idea is to represent every document as several distinct views and a co-training technique has been adopted for exploiting a huge set of unlabelled documents. Beginning from training sets of 10 per author, evaluating the effectiveness of co-training having limited labelled data. There are two methods and three views that are investigated—the logistic regression (LR) and the support vector machines (SVM) methods, and the character, the lexical, and the syntactic views. The results of the experiment show that the LR is much more effective for improving the AA co-training and its lexical view will perform best among three different views.

Ding et al. [11] made a proposal for the end-to-end digital investigation (EEDI) framework for a visualizable evidence-driven approach, called the VEA, aiming at facilitating the work of the cyber investigation. A comprehensive experiment and a stratified experiment in real-life Enron email datasets have demonstrated that this approach will achieve a higher accuracy than that of the traditional methods and this output is now easily visualized and also interpreted as traits of evidence. For identifying a plausible author for a text this approach will estimate this confidence for the result that is predicted based on a context of identification for linguistic evidence for all the candidates.

Layton et al. [12] made a proposal of a new method for creating a document profile or an author profile to detect features that are distinctive and this approach of re-entering will create more profiles that are accurate and are known as corpus of problems of authorship. This method is called re-entered local that determines the authorship by accurately using simple and ‘best matching author’ approach to the classification. By using a weighted voting scheme, the re-entered local profiles are shown to outperform the other methods in the AA, with an accuracy of about 69.9% in the ad-hoc and AA competition corpus,that represents a significant improvement in the related methods.

Ouamour and Sayoud [13] investigated a task of the AA on some old Arabic texts which were written by some ten ancient Arabic travellers and some features like the n-grams and the word n-grams have been used as the input of some sequential minimal optimization based SVM (SMO-SVM). The experiments of the AA, on the text database, have shown some interesting results having a classification precision of

80% and represents a rare-text mining work in Arabic and has shown some interesting points.

Brocardo et al. [14] made a proposal of authorship that is applied for some authentication which uses an online text based entry. The online document is decomposed into continuous blocks of short texts over the decisions of authentication and it has also been investigated of the texts that have 140, 280 and 500 characters. This feature set also includes traditional features like the lexical, the syntactic, the application specific features, and the new features that are extracted from the n-gram analysis. Furthermore, this approach includes a strategy for circumventing issues related to the unbalanced dataset, and also makes use of Information Gain and Mutual Information as a strategy of feature selection and the SVM used for classification.

Ouamour and Sayoud [15] made an investigation of the authorship of many short historical texts that have been written by ancient Arabic travellers ten in number and this is the Arabic dataset that has been collected by authors in the year 2011 that is known as the AAAT dataset. Many experiments of the AA have been conducted using these Arabic texts and using various lexical features like words, word-trig rams, word-big rams, rare words and word-trig rams. Also seven classifiers have been employed which are the Manhattan distance, the Cosine distance, the Stamatatos distance, the Canberra distance, the multi-layer perceptron (MLP), the SMO-SVM and the Linear Regression. For the task of evaluation many experiments of the AA have been conducted in the AAAT dataset using various quoted classifiers and features. The results have shown a good performance of attribution having an optimal score of 80%.

For training the features that are extracted from the poems of the Mukkoodar Pallu, the authors of other unknown poems may also be classified and the accuracy of classification by means of performing the classification on the datasets which contain features which have been extracted from various datasets which have been shown here in this work by making use of the C4.5 algorithm that is illustrated by Pandian et al. [16]. The actual results of the performing classification on datasets containing features from the datasets have been shown here in this work. The features like that of the sentences, numbers of characteristics and the accuracy of classification when the C4.5 algorithm has been used which has been illustrated. By means of doing this the authors of the other poems in the language of Tamil is identified and this can be helpful to the society as a whole.

Marukatat et al. [17] had presented a framework for identifying the Thai online messages and such the identification has been based on all the 53 writing attributes along with the chosen algorithms which are the SVM and also the C4.5 decision tree. The results of the experiment will indicate the accuracies that are completely achieved by the SVM along with the C4.5 that was 79 and 75%. The difference has not been significant statistically (at a confidence level of 95%) and for the identification of individual authors in certain cases the SVM is better than a C4.5. There have been other cases in which they are not distinguished from one author to another.

Wang [18] had tackled the task of classification in its author level, its article level, word level and sentence level having deep algorithms and non-deep algorithms with the GloVe word vectors that have been used as the word vectors that are pre-trained. Among such algorithms the recurrent neural network (RNN) in the sentence-level has got the best performance as it can capture the information and the word or sentence of the sequence information in the training dataset.

Homem and Carvalho [19] had considered the extraction of fingerprints form the texts which match them with the ones that are obtained from the authors. It further presents one more innovative algorithm of fuzzy fingerprint that was based on the fuzzy sets that are vector valued and the words along with the other features that are stylometric will be used for creating a fingerprint and its implementation has been based on the fast and compact algorithms that permit the method that has to be sued for the real time for large numbers of texts.

3 Methodology

Here in this section, the extraction features are used like the lexical, the syntactic, the structural and the ngram features. The C4.5, the Adaboost and the fuzzy classifiers have been described.

Lexical features The lexical features are connected to using the vocabulary of this language and this consists of breaking the text into a single atomic unit of the language called token. This can be a word or a character. In earlier studies a set of 100 frequent words are used for determining another documents' author and now more than 1000 frequent words are used for representing an author's style. The lexical features will encompass the frequency of the characters that are found in the text and the richness of the sentence, the line length the vocabulary, the n-grams and the lexical errors [20].

There are certain lexical features that measure the characters that include letters (both upper and lower case) digits and some special characteristics (like '@', '#', '$', '%', '(', ')', '{', '}', and so on.). The other lexical features are got by extracting the n-grams from a text and these are taken and formed by a contiguous sequence of n items. The N-grams will be a token that is formed using a contiguous sequence of the n items. A frequent n-grams will keep an important feature for stylistic reasons.

The richness of vocabulary will measure the diversity of the text by means of quantifying the number of unique vocabulary and the hapax legomenon (word occurring only once) and hapax dis legomenon (double or triple occurrences). The metric has been computed by means of dividing the total unique vocabulary using a total token number (each token is taken as a word.

Syntactic features The syntactic features are divided into an average punctuation and as a Part of Speech (POS). This is an unconscious trait that is more reliable and the punctuation here is an important rule that defines boundaries and identifies meaning by means of splitting the paragraph into sentences and further into different tokens. But this is not enough to analyse the punctuation of any document in words like 'Ph.D.' or 'uvic.ca' that include characters of punctuation as well. The Part-Of-Speech Tagging (POS tag or the POST) is categorizing tokens based on functions and the basic POS tags will include functional words expressing grammatical relationships like articles, adjectives, personal pronouns and auxiliary verbs [21].

The function words are those words that do not have a meaning on their own and are used for constructing sentences and many authors have many ways of structuring the words that are different from that of the others. The functions words like and, the and as are found in many documents and the number of times these words appear will identify anonymous documents [22].

There are also some other function words that are used in classification and they are not commonly used by the authors having special preferences and better grammar ability. Such words include words like "hence", "because", "whereas", "nevertheless", and "shall", that are substituted easily using other words. There are other syntactic features using punctuation marks like that of exclamation marks, question marks and semi-colons that are not used commonly in the documents. These symbols may be the preferences of certain authors and full stops and commas will show the style of some authors.

Structural features The structural features are used in learning as to how an individual organizes the document structure. The structural features were suggested first for the email in AA. In addition to this, the authors also made use of specific features to the emails like the presence and the absence of the greetings and the farewell remarks. Some of these structural features are [23]:

- The total number of lines
- The total number of sentences
- The total number of paragraphs
- The number of sentences in each paragraph
- The number of characters in each paragraph
- The number of words in each paragraph
- Having a greeting
- Having a separator between the paragraphs
- Using an e-mail as a signature
- Using the telephone as a signature
- Using the URL as a signature.

N-grams features The document that represents a feature vector may contain a single Boolean attribute for every work occurring within the documents and their training collection. On this method being generalized, by means of using the word sentences and for forming a sequence that is termed as the n-gram which is a feature. For the generation of such n-gram features, the n-gram features are discovered and this increases it for each n-gram with a list one n+1 gram which has the n-gram as its starting sequence. Therefore, an efficient algorithm which can generated feature sets by avoiding the generation of the n-grams and this algorithm will use the three parameters as the document collection, MaxGramSize and the MinFrequency. This Algorithm has been based on that of the APRIORI-algorithm which was for discovering the frequent item subsets in the databases. The final outcome of the project will be the learning algorithm that removes any stop word or a word sequence of a length of 2 or 3. The results showed an addition of the n-grams to that of the set of words and their representation that is used by the systems of text categorization which improves its performance and the sequences having a length of $n>3$ will not be useful and can bring down the performance

Single feature groups, that are combined have been described as below [24]:

The Character Bigrams (CBG): these provide robust indicators of authorship and various studies have confirmed them being superior in several large datasets.

The Character Trigrams (CTG): this captures a large amount of information that has an extra merit that represents common email acronyms like BTW, FAQ, FYI and so on.

The Word Unigrams (UNG): the frequency is one of the oldest and most reliable indicator of the performing of authorship in the n-gram features.

The Word Bigrams (WBG): this has been used in AA for long successfully.

The Word Trigrams (WTG): this is found to convey stylistic information as they are closer to the document's syntactic structure.

The Character n-grams approach phonology as well as morphology that captures some quantitative information relating to the syllable structure, the phonetics, the consonant clusters and the prefix and suffix structure. The word n-grams has an approach to syntax organization that includes lexical bundles, collocation structures and phrases. The Word n-grams are those sequences that are used excessively in various applications of NLP. The guessing of the next work will be an important subtask of recognition of speech, recognition of handwriting, spelling error detection and augmentative communication for those that are disabled, here the word identification becomes difficult as the input is quite noisy and therefore looking at previous words will give a clue as to what the subsequent ones will be [25].

Table 1 Growth in numbers of parameters for ngram models

Model	Parameters
2-Gram model	20,000 × 19,999 = 400 million
3-Gram model	$20{,}000^2 \times 19{,}999$ = 8 trillion
4-Gram model	$20{,}000^3 \times 19{,}999 = 1.6 \times 10^{17}$

The prediction ability is important in augmentative communication for the disabled. The using of n-gram models for representing the text will produce several parameters and if they are assumed conservatively that a corpus will contain a vocabulary of 20,000 words, there will be an estimate of various parameters as per Table 1. For extracting the word n-grams text is tokenized making the method language dependent as well as complicated.

An n-gram character is that sequence of the n adjacent characters. Given below is a sequence of seven characters in Japanese.

社長兼業務部長

As Japanese does not have any space between words we face an initial task of deciding the component words. This is in particular where the character sequence corresponds to two of the possible word sequences like "president, both, business, general manager" (= "a president and a general manager of the business") and "president, subsidiary business, Tsutomu (which is a name), general manager" (= ?). One needs some linguistic information for choosing the right alternative. The character level n-grams will not require taggers, tokenizers, or parsers or even NLP tools the n-gram extraction is a task that is language independent that makes this approach feasible for all problems of categorization.

These Character n-grams can also capture the stylistic information in the structural, syntactic or the lexical level. The character n-grams are very effective and various text collections in English, Chinese and Greek are used with good results. Also there is a variation method that achieved better results in ad hoc AA contest that is a collection of 13 text corpora in languages like English, Serbian Slavonic, Latin, Dutch, French and so on. The n-gram representation problem is the feature space size. The training set in used for an experiment that consists of 2500 small texts for a total of 12597 and 804 n-grams ranging between sizes of 2 and 11 (Table 2) which is an almost impossible feature space size that can be handled.

The decisions are made will be on the n-gram size for representing the problem and also the n-gram number that is used. The method of feature selection is applied for choosing important features that represent a class. The character n-grams of a certain length for the categorization of topics make use of SVM and the usual approach till today for the task of AA is to choose between either 3, or 4 or 5 g representations.

Table 2 Corpus n-grams

N	Number of n-grams
2	3000
3	26,767
4	115,488
5	315,553
6	655,697
7	1,127,202
8	1,691,381
9	2,298,435
10	2,899,521
11	3,464,760
Total	12,597,804

C4.5 classifier The Decision tree method is very effective in supervised learning aiming at the partition of the datasets into various groups that are as homogeneous as possible relating to the terms of the variable that is to be predicted. It further takes as an input a certain set of classified data and also outputs a tree which will resemble an orientation diagram in which every node (leaf) will denote a decision (a class) and each non-final node (internal) will represent a test. Every leaf represents a decision that belongs to a data class to verify test paths from root to the leaf [26]. This tree is quite simple and is technically easier to use. Sometimes, it is much more interesting to get a tree which is well adapted to the variables and their probabilities that are to be tested. Mostly the balanced tree will render good results. In case a sub tree leads to a situation that is unique all the sub trees will also get reduced to certain simple conclusions which simplify the process and not change the final result. At the time of construction of this decision tree it may be possible to handle data having certain attributes with an unknown value by means of evaluating the gain or the gain ratio for this attribute keeping in mind the records for which this attribute has been defined. By using this decision tree we can classify records having unknown values by means of estimating the outcomes and their probabilities. Its new criterion gain will be in the form below [27]:

$$Gain(p) = F(Info(T) - Info(p, T))$$

where:

$$Info(p, T) = \sum_{j=1}^{n} (p_j \times Entropie(p_j))$$
$$Info(T) = Entropy(T)$$

Here F = number of samples that are in the database with a known value for any given or total number of samples in that of a given set of attribute data.

The C4.5 will also manage the attributes with the values in the continuous intervals as: if C$_i$ attribute is a continuous interval of the values then it examines the attribute value in the training data. If these values are kept in ascending order, A$_1$, $A_2, \ldots, A_m$. In this case for every value that is partitioned between the records, those that have the values of C that is less than or equal to that of A$_j$ and the ones that have a larger A$_j$ value. For every partition a gain is calculated and the gain ration of the partition is maximised as gain is chosen [28]. The generation of a decision of the function best with that of a given training data set, will create a tree which will over-fit the data and will be sensitive to sample noise. These decision trees will not perform well in case of unseen samples. This will need pruning the tree for reducing the rate of error prediction. Pruning becomes a technique where machine learning reduces the tree size by means of removing those sections of the tree which provides some more power for classifying the instances. A dual pruning goal is the reduction of complexity of the final classifiers and also better the accuracy of prediction by reducing the overfitting and removal of sections of this classifier which is based on erroneous or noisy data.

This pruning algorithm has been based on a pessimistic estimate of the rate of error that is associated with a certain set of N cases, E that does not belong to the frequent classes. As opposed to E/N, C4.5 will determine the upper limit of the binomial probability in which E events are observed in N trials that use a specified confidence having a default value of 0.25. Pruning is carried out from the leaves to the root. This estimated error with the N cases and the E errors is the N times a pessimistic error. In case of a sub-tree that is replaced by that of a leaf when the latter is not higher than that of the former pruned sub-tree.

Adaptive boosting (Adaboost) classifier The Ada boost is an algorithm of machine learning that is used with various classifiers for improving the accuracy. The Ada boost is adaptive in the sense that the weak learners are now tweaked. The focus here is on the samples that are misclassified previously. In the initial stage all the samples have equal weight that may change in each of the boosting rounds. This is less susceptible to the problem of overfitting than that of the other learning algorithms. The individual learners are sometimes weak and for long as the performance is better the final model can converge to be a strong learner. The steps in Ada Boost classifiers are, Ada boost, Boosting, Bagging and Bootstrapping. The Boost classifier is in the form below [29]:

$$F_t(X) = \sum\nolimits_{t=1}^{T} f_t(x)$$

In which f_1 is the weak learner taking the value of the X as input and real value.

$$E_t = \sum\nolimits_i E[F_{t-1}(X_i) + \alpha_t h(X_i)]$$

In this, $F_{t-1}(X)$ is the boost classifier which built the previous training stage.

This Ada Boost algorithm, that was introduced in the year 1995 by Freund and Schapire, had solved many practical difficulties of the boosting algorithms will be the focus in this paper. The Pseudo code for the Ada Boost has been given below in the generalized form that has been given by Singer and Schapire. This algorithm takes as an input the training set $(x_1, y_1), \ldots (x_m, y_m)$ in which each x_i will belong to any domain or an instance space X and every label y_i is in a certain label set Y. For most part of this paper it has been assumed that $Y = \{-1, +1\}$. The Ada Boost calls a base learning algorithm that is weak continuously for a series of three rounds $t = 1, \ldots, T$. The main idea of this algorithm has to maintain a distribution or a set of weights in the training set. The distribution's weight on the training example I, on round t has been denoted by $D_t(i)$. Initially, all of these weights have been set equally, but on each of the rounds the weights of all the wrongly classified examples that are increased to make sure that this base learner has been the focus of hard examples in various training sets [30].

```
Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \epsilon X, y_i \epsilon = \{-1, +1\}$.
Intialize $D_1 = 1/m$.
For $t = 1, \ldots, T$:
Train base learner $u \sin g$ distribution $D_t$.
Get base classifier $h_t : X \to R$
Choose $\alpha_t \epsilon R$.
UPdate:
$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t, Y_i h_t(x_i))}{Z_t}$
Where $Z - t$ is a normalization factor
(Chosen so that $D_{t+1}$
will be a distribution).
output the final classifier:
$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$
```

Fuzzy classifier The Fuzzy classification is that process of grouping of elements that are into a fuzzy set. In this fuzzy classification technique there is a membership function μ which will indicate if an individual belongs to a member of a certain class. This class is that set which is defined using a specific property and all the objects that have such property are the elements belonging to that particular class. The process of classification will evaluate the set of objects and checks if they will accomplish the property and its classification. In case there is a match then it will belong to a member of the corresponding class [31]. In this the fuzzy classifier will be used for classifying the given input text for every author and extract features specific to each of them. Every author has a set of features that are unique to him and if it gives a new text or wants to identify an unknown text's author then this author will be the most likely to be assigned.

A fuzzifier will converts any crisp value within the membership degree applying such membership functions that determine the crisp value association certainty for a particular linguistic value. For a temperature membership function, a value of temperature of – 2°C has been classified as about 20% freezing and about 80% Cold. These membership functions may have various shapes the frequently used being the triangular, the trapezoidal, and the Gaussian shaped. These membership functions have been defined by relying on the domain knowledge or by an application of techniques of learning like the neural networks and the genetic algorithms. For any of the set X, there is a membership function on the X which is a function from that of X to a real unit interval [0, 1]. This membership function that represents one fuzzy set that is normally denoted by μ_A. For every element x of X, its value μ_A(x) will be called the degree of membership of x in its fuzzy set. A membership degree μ_A(x) will quantify a membership grade for an element x to that of the fuzzy set. A value 0 will mean that the x is not the member of this fuzzy set; a value 1 is the x that is fully a member of any fuzzy set. The values that exist between 0 and 1 will characterize the fuzzy members, belonging to a fuzzy set partially. The rules of Fuzzy classification will contain fuzzy sets in the antecedent and a class label that is in the consequent. If we denote the data set that had D data points and n variables as Z = [X y], in which the input matrix X and the output vector y have been given as [32]:

$$x = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{1,2} & \cdots & x_{2,n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{D,1} & x_{D,2} & \cdots & x_{D,n} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_D \end{bmatrix}$$

The fuzzy classification is performed as below:

$$R_i = IFx_1\, is\, A_{i,1} \ldots and\, x_n\, is\, A_{i,n}\, then\, g_i\; i = 1, \ldots R,$$

In which R is the actual number of rules, $A_{ij},\ j = 1, \ldots, n$ is the membership function, $g_i \epsilon \{1, \ldots, C\}$ will be the rule that is consequent and C will be the number of classes in the data set. For each of the data points the fulfilment degree of a rule has been computed as:

$$\beta_i(X_k) = \prod_{j=1}^{n} A_{i,j}(x_{k,j})$$

This rule has a high degree of fulfilment and has been declared as the winner rule (where the winner will take all the strategy).

The rules will be in a format—'If error is Ai, and the change in the error Bi then the output will be Ci'. And the if 'part' of any rule will be called a rule-antecedent and will be a description of one process state relating to a logical combination of the various atomic fuzzy propositions. Now the 'then' part of this rule is known as a rule consequent and will be a description of the total control output for the logical combinations of such fuzzy propositions.

The classifier is the rule that is consequent to the rule associated. There are certain other types of fuzzy rules and the t-norms that are applied to the reasoning and the fuzzy classifier property that are discussed in detail. The accuracy of the fuzzy classifiers has been measured using various miscalculations. But there is no generic way to interpret this. Most often the interpretability has been measured using the number of total antecedents in the rules (the total strength of rule). It has been mentioned that the number of rules with the total length of rules can lead to preventions of overfitting. As a consequence, it has been beneficial for being used by both the objectives. These objectives are to be minimised based on the misclassifications and the rules and the total strength of the rules.

4 Results and discussion

Dataset were obtained from Amazon reviewers. A total of 5 authors with 40 review each were selected. Using tenfold cross validation the dataset were evaluated. Sample amazon book: Batman R.I.P. (Grant Morrison and Tony Daniel), Bone Crossed (Patricia Briggs), Cat Playing Cupid (Shirley Rousseau Murphy), Dream Warrior (Sherrilyn Kenyon), Great Powers of America (Thomas P.M. Barnett), Maelstrom (Taylor Anderson), Promises in Death (J.D. Robb), Revelation (C. J. Sansom) and Run for Your Life (James Patterson and Michael Ledwidge).

Experimental setup: Experiments were carried out with 2 authors, 3 authors and 5 authors. Experiments were conducted using syntactic + Structural and proposed syntactic + structural + n gram features. In this section, the C4.5, random tree, fuzzy classifier and Adaboost techniques are evaluated using syntactic, structural and ngram features. The accuracy in two, three and five authors are shown in Table 3 and Figs. 1, 2, and 3.

Table 3 Percentage improvement for two, three and five authors

Classifiers used	Syntactic + structural compared with syntactic + structural + n gram		
	Two authors	Three authors	Five authors
C4.5	4.95%	9.18%	8.39%
Random Tree	5.97%	2.13%	4.53%
Fuzzy Classifier	6.63%	4.04%	3.68%
Ada Boost	7.97%	3.92%	3.57%

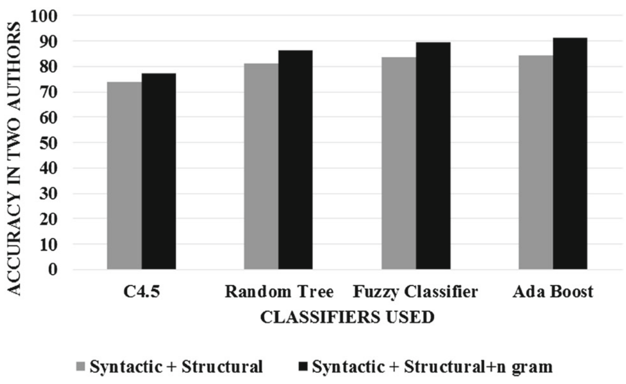

Fig. 1 Accuracy in two authors

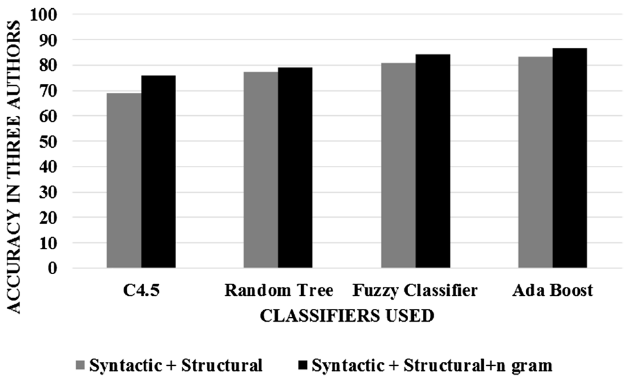

Fig. 2 Accuracy in three authors

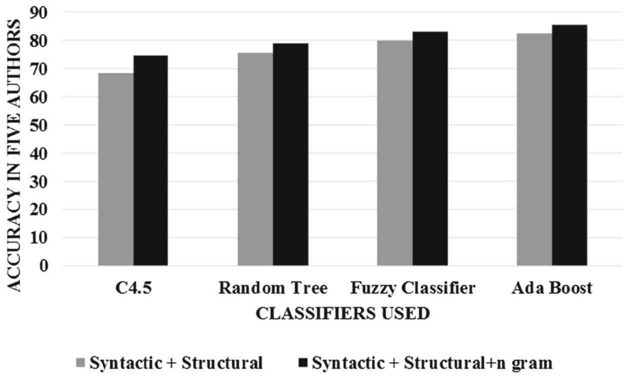

Fig. 3 Accuracy in five authors

From the Fig. 1, it can be observed that the syntactic + structural + ngram has higher accuracy in two authors by 4.95% for C4.5, by 5.97% for random tree, by 6.63% for fuzzy classifier and by 7.97% for Adaboost when compared with syntactic + structural.

From the Fig. 2, it can be observed that the syntactic + structural + ngram has higher accuracy in three authors by 9.18% for C4.5, by 2.13% for random tree, by 4.04% for fuzzy classifier and by 3.92% for Adaboost when compared with syntactic + structural.

From the Fig. 3, it can be observed that the syntactic + structural + ngram has higher accuracy in five authors by 8.39% for C4.5, by 4.53% for random tree, by 3.68% for fuzzy classifier and by 3.57% for Adaboost when compared with syntactic + structural.

5 Conclusion

Here in this work, the authorship identification framework of the different online messages has been proposed. In this work, a framework for authorship identification of online messages is proposed. For the purpose of evaluation of effectiveness of this framework several experiments have beelen conducted based on the accuracy in either two, three or five authors. The results have shown that the syntactic + structural + n-gram will have a much higher accuracy in both of the authors by about 4.95% for the C4.5, by about 5.97% for the random tree, by about 6.63% for the fuzzy classifier and by about 7.97% for the Ada boost when a comparison is made with that of the syntactic + structural. This syntactic + structural + n-gram has proved to have a higher level of accuracy in three of the authors by about 9.18% for the C4.5, by about 2.13% for the random tree, by about 4.04% for the fuzzy classifier and by about 3.92% for the Ada-boost when this has been compared with that of the syntactic + structural. This syntactic + structural + n-gram has a much higher accuracy in the five authors by about 8.39% for the C4.5, by about 4.53% for the random tree, by about 3.68% for the fuzzy classifier and by about 3.57% for the Ada boost when is has been compared with that of the syntactic + structural. For all future work, the plan will be a relationship analysis among the features that are used for the process of extraction, the optimal two word phrases and the modification of the learning engine for improving further the performance of classification for the forensics.

References

1. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
2. Iqbal, F., Binsalleeh, H., Fung, B.C., Debbabi, M.: A unified data mining solution for authorship analysis in anonymous textual communications. Inf. Sci. **231**, 98–112 (2013)
3. Zhang, C., Wu, X., Niu, Z., Ding, W.: Authorship identification from unstructured texts. Knowl. Syst. **66**, 99–111 (2014)
4. Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C.A., Chen, H.: Evaluating text visualization for authorship analysis. Secur. Inf. **3**(1), 1 (2014)
5. Cristani, M., Roffo, G., Segalin, C., Bazzani, L., Vinciarelli, A., Murino, V.: Conversationally-inspired stylometric features for authorship attribution in instant messaging. In: Proceedings of the

20th ACM International Conference on Multimedia, pp. 1121–1124. ACM (2012)

6. Nirkhi, S., Dharaskar, R.V.: Authorship identification in digital forensics using machine learning approach. Int. J. Latest Trends Eng. Technol. (IJLTET) **5**(1) (2015)
7. Nirkhi, S., Dharaskar, R.V.: Comparative study of authorship identification techniques for cyber forensics analysis. arXiv:1401.6118 (2013)
8. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship verification via k-nearest neighbor estimation. Notebook for pan at CLEF (2013)
9. Zamani, H., Esfahani, H.N., Babaie, P., Abnar, S., Dehghani, M., Shakery, A.: . Authorship identification using dynamic selection of features from probabilistic feature set. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 128–140. Springer, New York (2014)
10. Qian, T., Liu, B., Zhong, M., He, G.: Co-training on authorship attribution with very fewlabeled examples: methods vs. views. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 903–906. ACM (2014)
11. Ding, S.H., Fung, B., Debbabi, M.: A visualizable evidence-driven approach for authorship attribution. ACM Trans. Inf. Syst. Secur. **17**(3), 12 (2015)
12. Layton, R., Watters, P., Dazeley, R.: Recentred local profiles for authorship attribution. Nat. Lang. Eng. **18**(03), 293–312 (2012)
13. Ouamour, S., Sayoud, H.: Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier. In: IEEE International Conference on Communications and Information Technology (ICCIT), pp. 44–47 (2012)
14. Brocardo, M.L., Traore, I., Woungang, I.: Authorship verification of e-mail and tweet messages applied for continuous authentication. J. Comput. Syst. Sci. **81**(8), 1429–1440 (2015)
15. Ouamour, S., Sayoud, H.: Authorship attribution of short historical arabic texts based on lexical features. In: IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 144–147 (2013)
16. Pandian, A., Ramalingam, V.V., Preet, R.V.: Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) Using C4. 5 Algorithm. Indian J. Sci. Technol. **9**(47) (2016)
17. Marukatat, R., Somkiadcharoen, R., Nalintasnai, R., Aramboonpong, T.: Authorship attribution analysis of Thai online messages. In: IEEE International Conference on Information Science and Applications (ICISA), pp. 1–4 (2014)
18. Wang, L.Z.: News authorship identification with deep learning (2017)
19. Homem, N., Carvalho, J.P.: Authorship identification and author fuzzy "fingerprints". In: Annual Meeting of the North American IEEE Fuzzy Information Processing Society (NAFIPS), pp. 1–6 (2011)
20. Luiz Brocardo, M., Traore, I., Saad, S., Woungang, I.: Verifying online user identity using stylometric analysis for short messages. J. Netw. **9**(12), 3347–3355 (2014)
21. Brocardo, M.L., Traore, I., Woungang, I.: Toward a framework for continuous authentication using stylometry. In: IEEE 28th International Conference on Advanced Information Networking and Applications, pp. 106–115 (2014)
22. Tan, R.H.R., Tsai, F.S.: Authorship identification for online text. In: IEEE International Conference on Cyberworlds (CW), pp. 155–162 (2010)
23. El, S.E.M., Kassou, I.: Authorship analysis studies: a survey. Int. J. Comput. Appl. **86**(12) (2014)
24. Mikros, G.K., Perifanos, K.: Authorship identification in large email collections: experiments using features that belong to different linguistic levels. Notebook for PAN at CLEF (2011)
25. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 77–86. Springer, Berlin (2006)
26. Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M.: A comparative study of decision tree ID3 and C4. 5. Int. J. Adv. Comput. Sci. Appl. **4**(2) (2014)
27. Sharma, S., Agrawal, J., Sharma, S.: Classification through machine learning technique: C4. 5 algorithm based on various entropies. Int. J. Comput. Appl. **82**(16) (2013)
28. Cintra, M.E., Monard, M.C., Camargo, H.A.: A fuzzy decision tree algorithm based on C4.5. Mathw. Soft Comput. **20**, 56–62 (2013)
29. Kaur, E.N., Kaur, E.Y.: Object classification Techniques using Machine Learning Model. Int. J. Comput. Trends Technol. **18**(4) (2014)
30. Schapire, R.E.: The boosting approach to machine learning: an overview. In: Nonlinear Estimation and Classification, pp. 149–171. Springer, New York (2003)
31. Pulkkinen, P., Koivisto, H.: Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms. Int. J. Approx. Reason. **48**(2), 526–543 (2008)
32. Elayidom, M.S., Jose, C., Puthussery, A., Sasi, N.K.: Text classification for authorship attribution analysis. arXiv:1310.4909 (2013)

L. Srinivasan is working as an Assistant Professor at Dhirajlal Gandhi College of Technology, Salem, Tamilnadu. He Received his M.Tech(IT) from Anna University Coimbatore and B.E(IT) from V.L.B Janakiammal College of Engineering & Technology, Coimbatore in 2011 and 2004, respectively. His current research includes Text mining,big data analytics. He has published more than 10 research articles in leading journals, conference proceedings. He has more than twelve years of professional experience in Various Engineering Colleges in Tamilnadu, India.

C. Nalini is working as a Professor at Department of Information Technology, Kongu Engineering College, Erode, India. She has 25 years of teaching experience. Her current research interest includes data mining, big data analytics, ubiquitous computing, privacy preserving, etc.