Approaches of Authorship Verification

Shaimaa Ayman
Department of Computer Science
Zagazig University, Sharqiyah, Egypt

Mohamed Eisa
Department of Technology and Information System
Port Said University, Egypt

Fifi Farouk
Department of Technology and Information System
Port Said University, Egypt

ABSTRACT

Nowadays, there are massive amounts of texts in digital form in digital libraries, online journalism, and social networks; for example, Twitter is estimated half a billion tweets are sent out each day. The expanded usage of Online Social Network (OSN) has become necessary to appear to grow of Authorship Verification (AV), OSN is the environment in which users can connect with other users to discuss ideas of any topics then expansion data and information. AV considered as a resource of researches and information in different ways, as is the case Sentiment Analysis (SA). Information that gained from Twitter and Facebook or any other OSN is considered valuable in some areas such as public opinion organizations and online marketing. The crimes also increased over on the internet with textual data. To reduce the problems raised on text through the internet, the researchers have attracted to authorship analysis which is one of the important areas. AV is a type of authorship analysis that is used to verify an author by checking whether the text document is written by the disputed author. The accuracy of AV depends primarily on the features used to distinguish the writing style of documents. In previous works of AV, researchers proposed several types of stylistic features for distinguishing the writing style of the authors. The researchers analyzed that the AV performance was weak when used stylistic features alone in the experiments. Therefore, researchers resorted to more accurate methods that compute the features by using the weight measures. The weight measures calculate the document weights of training, and test documents. Then, the competition between the weights of training document and the weights of test document were implemented; to verify the author of the document.

Keywords

Online Social Networks, Authorship Analysis, Authorship Verification, Stylistic Features, Term Weight Measures.

1. INTRODUCTION

With the increasing and proliferation in digital information, Online Social Network (OSN) and applications such as security [1,2] and plagiarism detection [3] aim to increase the need for reliable Authorship Verification (AV) techniques. AV is one subfield of authorship analysis. The task of AV is to estimate whether the text in dispute was written by the same author of the known texts. AV techniques are depending on confidence measurements. AV may be known as a subtask for the authorship attribution when it is appropriate for solving a binary problem: whether or not the text belongs to a given author. On the other hand, because of its usage in humanitarian and forensic science, it is also seen as an autonomous task: resolving conflicts over copyright, and recognizing multiple pseudonyms of the same person. In the context of OSN, a digital document can be used as evidence to confirm that a suspect is a criminal if he or she is the author of the document or not. The authorship verification research may be used both a one-class classification approach [4, 5], and a two-class classification approach [6] to solve the authorship verification task. The idea of one-class is to make the most of known texts given and set a rule of classification to determine the label of unknown text. To solve the search problem as a two-class classification problem, the collection and preparation of data are of great importance. As recently as the 1990s, through social media found a huge amount of electronic texts, that have been needed to handle information. The AV is an open-set problem to determine the actual author from a lot of documents. The question is; if an anonymous document was written by a candidate author or not. The answers will be positive for the true author and a negative for all the others. AV problem is a specific task of the authorship attribution with an open set of candidate authors. Koppel and Schler [4] clarify the “unmasking” method, this method requires that the input documents be very long they chunk each document to equal sections (500 words) without decomposing paragraphs and using machine learning algorithms to distinguish them. This chunk of text must belong to present the statistical representativeness. Authorship verification has been affected with developed of more fields such as machine learning research, information retrieval research, and Natural Language Processing (NLP) research. Features representation is considered as the main problem in text classification. More studies have been a focus on the bag-of-words, bigrams, unigrams, and N-grams models for feature extraction.

2. DATA SAMPLING Of AUTHORSHIP VERIFICATION APPROACHS

Every problem of authorship identification in general consists of a set of candidate authors, a set of text samples of known authors that include all of the candidate authors (training set), and a set of text samples from unknown authors (test set). Data sampling can be applied by using one of the four main data sampling approaches: a profile-based approach, an instance-based approach, one-class Classification, and two-class Classification.

Authorship attribution can be distinguished according to whether they deal with each training text individual or cumulative for each author. There are some links for all the training texts available to each author in one large file, and extract a cumulative representation of that author's style usually called the author's profile [7]. On the other hand, another approach needs various training text samples per author to improve the accuracy of the attribution model. Each training text is represented separately as a separate instance of author style [8].

2.1 Profile-Based Approach

It is a method of integrating all author's training texts and creating an author's profile. Each author's features are extracted out from the concatenated text. In the AV, extracted characteristics are used to determine the most likely author of the text in dispute. However, a profile-based approach is criticized for wasting a lot of information due to the process of creating profile-based features needed to delete all of the contents that dissimilar of the same author. Profile-based approach provide a simple training process.

The training process only requires profile extraction for the candidate authors. Then, the attribution model is typically based on a distance function, which measures the differences between the profile of an unknown text and each author's profile [9]. A typical profile-based approach architecture is shown in Figure 1. Stamatatos [10] proposed an intrinsic AV approach that reshaped the original method [11]. In [12], a profile-based authorship attribution approach is proposed for Chinese online messages. Used N-grams techniques to extract frequent sequences from extensive linguistic elements including Chinese characters, English characters, digits, symbols. Developed a profile-based approach to represent the suspects to category profiles. Developed a frequent sequence standard way to solve the class imbalance problem. V Kešelj [11] used a distance dissimilarity measure to determine the differences between the constructed profiles, based on the most frequent N-grams of the text. Applied probabilistic modeling in the Federalist Papers [13] on authorship recognition to classify the author by taking the maximum conditional likelihood between the author's texts and the anonymous document.

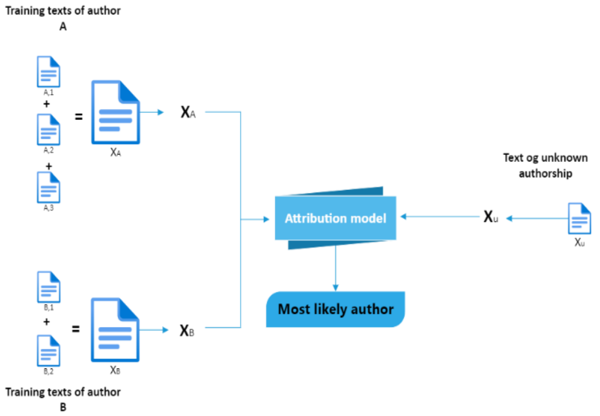

Figure 1: Profile-based Approach Architecture [9]

2.2 Instance-Based Approach

An instance-based approach, which is used in more of the recent authorship attribution research; it can retain most of the information from the given texts, and extracted features are applied to a machine learning classifier. Every training document is handled individually, and contributes as an instance in the identification model for authorships. A vector of attributes is represented in each text sample of the training corpus, and a classification algorithm is trained to build an attribution model using a set of instances of known authorship (training set). A machine learning algorithm takes vectors from the numerical features that describe a specific author's texts and creates a corresponding attribution model that is used to identify the possible author of an anonymous document. The model will then be able to estimate an unknown text of the true author [9]. An instance-based approach architecture is shown in Figure 2.

S.Argamon [14], takes the advantages of multiplicative orthographic learning to distinguish authors within a newsgroup corpus. Koppel [15], relys primarily on misspelling features to identify the author in the email text; in addition to other lexical and syntactic collections. De Vel [16], analyzed stylistics attributes used Support Vector Machines (SVM) to detect plagiarism in e-mail corpus, SVM is the best of the machine learning algorithms in this field [17].

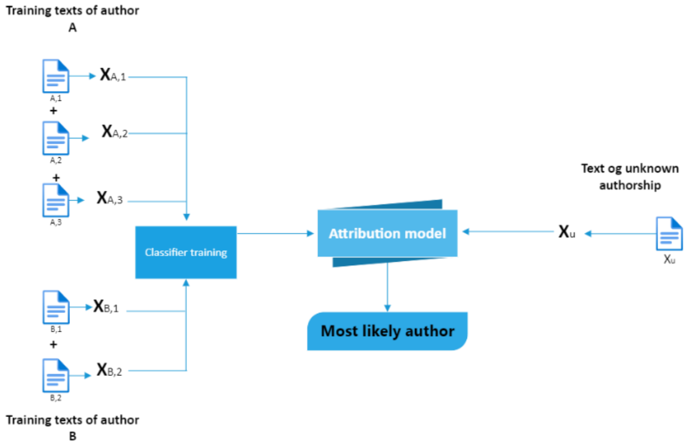

Figure 2: Instance-Based Approach Architecture [9]

2.3 One-class Classification Approach

Is a field of machine learning that provides techniques for outlier and anomaly detection. Classification can be explained by one category where there is only one class. Thus, the result is simply that the studied object is in class or not. The one-class classification description appears in Figure 3. Koppel[4], ignored negative examples, and treat with AV as a true one-class classification problem. AV research features strongly with the one-class classification characteristics. Magdalena [18] proposed AV using a proximity-based method for a one-class classification that applies the Common N-Gram (CNG) dissimilarity measure.

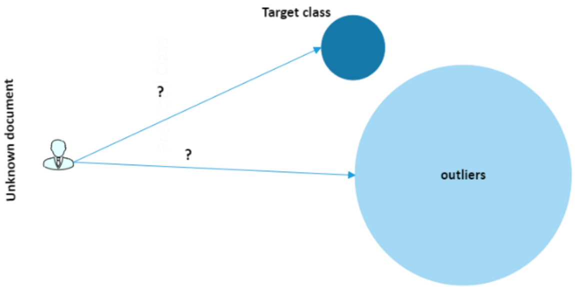

Figure 3: One-Class Classification [19]

2.4 Two-class Classification Approach

The one-class approach can be very unfair when the author's texts are limited, and therefore the result is not accurate. In this case, for machine learning classifiers the outlier class can be generated to learn and distinguish between the two classes. Two-class problems contain information for all classes, and also allows monitoring of category errors; including false positives. Figure 4 describes the two-class classification with the appropriate outlier selection.

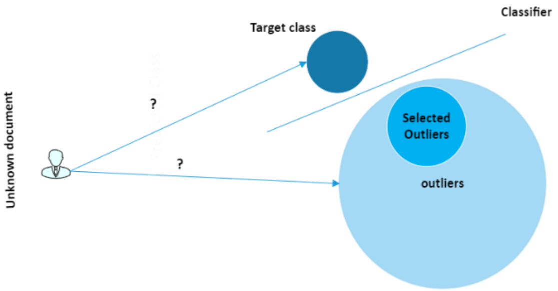

Figure 4: Two-class classification with Appropriately Selected Outliers [19]

Figure 5 describes the two-class classification with inappropriately selected outliers. This shows that if the selected outliers are distant from the target class such as Selected Outlier A, then every classification between Classifier One and Classifier Two in Figure 5 is considered to be successful in classifying between the target class and the Selected Outlier A. And hence, the model built on the basis of Selected Outlier A is more likely to mark other outliers similar to the target class, such as all Selected Outlier B instances, as target class. Hence misclassification is more likely between target class and outliers.

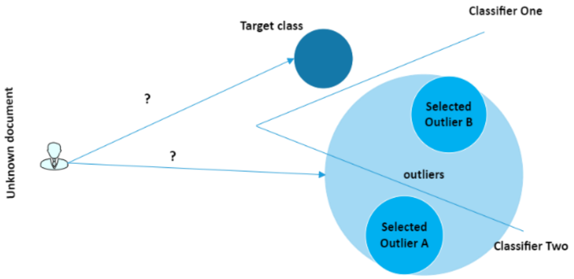

Figure 5: Two-class classification with inappropriately selected outliers [19]

3. EXISTING AUTHORSHIP VERIFICATION APPROACHES

Marcelo Luiz Brocardo et al [20] have studied the effect of the possibility of using stylometry for AV for the short online message. Based on the combination of supervised learning and N-grams analysis, more specifically "a combination of logistic regression and SVM (so-called SVM-LR method)". They have estimated their experimental approach by using the Enron emails dataset.

Oren Halvani [21] have clarified AV via k-Nearest Neighbor (k-NN) estimation notebook for PAN dataset, they depend on the k-NN classifier in their verification algorithm, it was working on numeric values only. This method is based on the combination of suitable feature categories. For each chosen feature category applying a k-NN classifier to calculate a style deviation score between the training documents of the true author and the document from an author, who claims to be true author, depending on the score and a given threshold.

Koppel et al [22] have designed Unmasking algorithm to measure the degradation rate of accuracy of the learned models, as an iterative process by eliminating the most different features, they suppose that; when the accuracy is more degraded, that because the author of the test document matches with the training author. Koppel divide the document into some sets of writings, where each chunk in each set contains at least 500 words, without decomposing the paragraphs. It is very suitable for the corpus of the books.

Smita Nirkhi et al [23] have explored AV of online messages as a clustering problem, and used unsupervised machine learning methods. To solve the AV problem used cluster analysis and multidimensional scaling techniques; which provides visualization of clusters, helpful to the investigator to visualize the results. Cluster analysis: used clustering to show the similarity between two documents. Hierarchical agglomerative clustering is used the clustering algorithm. There are small pairs of closely related documents are combined and form groups. After that these small groups are combined into the larger group until all the documents are connected into a single large cluster. The text documents written by one author are similar and are placed on neighboring branches. Multi-Dimensional Scaling (MDS) is visualization technique based on a distance matrix. MDS represents the original matrix by a two-dimensional map where the values in the vector become co-ordinates of the document. This representation is more similar documents appear closer together.

Benedikt et al [24] have proposed a new algorithm to compare forensic texts called ADHOMINEM, where described by an attention-based Siamese network topology, which is learning linguistically analysis features based on the visualization of the internal attention weights such as non-standard lexical forms, spelling errors, and expressions that change in other styles of the standard.

4. FEATURES TYPES OF AUTHORSHIP ANALYSIS

In the study of authorship analysis by general, the most traditional features are stylometric features, while some researchers such as Lambers and Veenman [25] have tried to use compression distances between texts as a new-fashioned feature to categorized the problem of AV. So, both stylometric features and compression distance features will be discussed in this section.

4.1 Stylometric Features

Stylometry is a branch of computational linguistics that studies quantitative estimates of linguistic features in the Natural Language Processing (NLP). To apply machine learning techniques to stylometry measurement, some Python libraries are used to provide the basis for statistical analysis of text data, the Natural Language Tools Library (NLTK) [26]. Chen and Hao’s [27] have used 150 stylometric features for applying authorship similarity detection from e-mail messages that using 40 authors of the dataset of Enron. The accuracy rates for 10 and 15 short e-mails were 84% and 89% respectively. The number and length of emails have impacted the final performance for several cases. The best result achieved when used SVM and decision tree as basic methods with the increasing length of e-mails, the performance of PCA and K-means clustering outweighed in this research for all cases. When deciding the writing types, the stylometric features of the documents are of considerable importance. The most famous of stylometric features; lexical features, character features, syntactic features, semantic features, and application-specific features. Will be described below.

4.1.1 Lexical Features

The lexical feature is a simple way to represent text, known as token-based features or word-based features. Lexical feature considered as a language-independent, meaning that they can be applied with the aid of a tokenizer to all languages. Darnes [28] used lexical-syntactic with graph-based features to represent the unique writing style of a given author, the runtime was large but the performance was good. It was good in the Spanish language. Iqbal and et [29] have used some important lexical features like the richness of vocabulary, word length distribution, and the average number of words. Some researchers have used word N-grams to solve the authorship attribution problems. However, the richness of vocabulary is might be ineffective because a great many word types from the texts are hapax legomena, meaning they appear once in the whole text. So, Hoover [30] proposed that the difference between texts written by the same author can be different as the texts written by different authors. Stamatatos [31] have used preprocessing procedures for authorship attribution that using text distortion, character N-grams, and word N-grams.

4.1.2 Character Features

According to text measures; a text is a sequence of characters, so a lot of character measures can be defined like alphabetic characters count, uppercase and lowercase characters count, digit characters count, punctuation marks count, letter frequencies, and the total number of characters per token. Extract frequencies of N-grams on the character, it can capture nuances of style, including lexical information hints of contextual information use of capitalization and punctuation. When the question texts are noisy, it's able to tolerate it and containing grammatical errors. Note that, such errors could be considered as the author's traits in the style-based categorization of text. This detail is also captured by N-grams of character like in [15]. Brocardo et al. [20] studied the effect of the possibility of using stylometry for AV for the short online message. Based on the combination of supervised learning and N-grams analysis. They used the Enron emails dataset including 500 characters of block size for 87 authors. They used stylometric techniques through linguistic analysis and writing styles. They evaluate the performance of their approach through a 10-fold validation test. The Equal Error Rate (EER) was 14.35%. There are some limitations in their model used one type of features and not good also to handle short message content 10 to 50 characters like Twitter. Performed Grieve[32] research to evaluate 39 textual measurement techniques including word-length, word frequency, sentence length, graph frequency, vocabulary richness, punctuation mark frequency, collocation frequency features, and character-level N-grams frequency features. Grieve deals with 1600 texts with average text length 937 words in the range from 500 to 2000 words from 40 authors with similar backgrounds. Found out that the results on the corpus were from word and punctuation mark combination, character 2-grams/bigrams, and 3-grams/trigrams.

4.1.3 Syntactic Features

For syntactic features method is to use syntactic information, the idea is that the authors tend to employ similar syntactic patterns unconsciously. Thus, in contrast with lexical information, syntactic information is considered more of a credible authorial fingerprint. Besides, the benefit of function words in style representation demonstrates the utility of syntactic information as they are normally. Such information includes robust and accurate NLP tools capable of performing syntactic text analysis. This reality means that extraction of the syntactic measure is a language-dependent process since it depends on the availability of a parser capable of processing a specific natural language with reasonably high precision. English part-of-speech tagging is one of the popular representations. Many researchers like Koppel, Schler, and Argamon [33] have also adopted the frequencies of part-of-speech tagging as a deterministic stylometric feature.

4.1.4 Semantic Features

Compared with poor features such as character N-grams; semantic features are called rich stylometric features [34]. By now, it should be clear that the more detailed the text analysis needed to extract stylometric features, the less accurate and noisier measures produced. Simply depending on the rich stylometric features; the outcome of Tanguy [34] did not achieve satisfactory results, but the combination of rich features with poor features has improved the results obtained by using them separately. WordNet, a Princeton University project, is a top-quality source of word synonyms and hypernyms proposed by Fellbaum [35]. NLP tools can be implemented successfully to low-level tasks; such as text chunking, sentence splitting, POS tagging, and partial parsing. Argamon et al. [36] have perhaps defined the most effective method of exploiting semantic knowledge so far. Inspired by the Systemic Functional Grammar Theory MAK Halliday [37], they defined a set of functional features that combine certain words or phrases with semantic information. Another approach to extract semantic features described by McCarthy et al [38], based on WordNet [35] estimated information on synonyms and word hypernyms and identified causal verbs.

4.1.5 Application-Specific Features

The lexical, character, syntactic, and semantic features identified before are application-independent. They can be extracted from any textual data given the availability of the suitable NLP tools, and training required for their measurement. The person may identify application-specific measures to better represent style differences in a given text field. There are a lot of types to application-specific measures like Functional, Structural, Content-specific, and Language-specific. Structural measures include the use of greetings and farewells in the messages, paragraph length, types of signatures, and use of indentation proposed by Zheng et al [17]. Content-specific keywords may be used to capture the best properties of an author's style within a given text-domain. More precisely, given that the texts in question deal with similar subjects and are of the same genre, it is important to describe similar terms that are frequently used within that subject or genre. For example, in the sense of newsgroup analysis for online messages proposed by Zhang [39]. Stylometric features try to avoid content-specific information to be more authoritative in cross-topic texts. However, in cases where all texts available to all nominee authors are in the same subject area, carefully selected content-based information may reveal some of the author's options. The content keywords used may be better for capturing author style characteristics within a specific text range.

4.2 Compression Distance Features

Compression distance features can be applied to both profile-based approach and instance-based approach, while Stamatatos [9] argues that the use of compression features in comparison with the profile-based approach is more effective based on a study of the literature concerned. As for the compression versions; the choice of the most suitable compression algorithm and the calculation of the compression distance is of high importance.

The most successful of the compression-based approaches follow the profile-based methodology Marton [40]. A compression-based is dissimilarity method which is based on Kolmogorov complexity proposed by Keogh [41]. The

Compression-based Dissimilarity Method (CDM) is defined in equation (1) as follows:

$$CDM(x,y) = \frac{C(xy)}{C(x)+C(y)} \qquad (1)$$

Where C is the compression algorithm, $C(x)$ is the compressed length of the compressed document) of object x, $C(y)$ is the compressed length of object y, and $C(xy)$ is the compressed size of the concatenated object xy. When x and y are the same, then CDM (x, y) is close to 0.5, and when x and y are completely different, then CDM (x, y) is approaching 1.

The compression-based dissimilarity approach developed jointly by Li et al [42] and Cilibrasi et al [43] is called the Normalized Compression Distance (NCD). The definition of the (NCD) is shown in equation (2):

$$NCD(x,y) = \frac{C(xy) - \min\{C(x),C(y)\}}{\max\{C(x),C(y)\}} \qquad (2)$$

Where $C(xy)$ is the compressed size of the concatenated object xy, $C(x)$ is the compressed result of object x, and $C(y)$ is the compressed size of object$y$.

Chen, Li, and their partners have developed the compression method called the Chen-Li Metric (CLM), Li [44] and Chen [45]; the formulation is as follows in eq (3):

$$CLM(x,y) = 1 - \frac{C(x) - C(x|y)}{C(xy)} \qquad (3)$$

Where x and y represent the objects that are to be compressed by algorithm C, and$C(x|y) = C(xy) - C(y)$. The result of this metric is between 0 and 1. When x and y are completely the same, CLM (x, y) is 0, and when x and y are completely different, CLM (x, y) is 1.

Another new compression measure was developed based on cosine-vector dissimilarity measure by Sculley et al [46]. The equation of the Compression-based Cosine (CosS) metric is as follows in eq (4):

$$CosS(x,y) = 1 - \frac{C(x) + C(y) - C(xy)}{\sqrt{C(x)C(y)}} \qquad (4)$$

5. FEATURE SELECTION AND EXTRACTION

Some types of features such as lexical and character features can significantly increase the dimensionality of the set of features. In such cases, it is possible to apply feature selection algorithms to reduce the representation's dimensionality proposed by Forman [47]. Thus, it allows the classification algorithm to avoid overfitting on the training data.

Another approach to reducing dimensionality is via feature extraction proposed by Sebastiani [48] combining the original collection of features produces a new collection of "synthetic" features. For authorship research studies, the most conventional feature-extraction technique is the main component research which provides linear combinations of the initial features. The two most important key components can then be used in a two-dimensional space to represent the texts.

6. COMPUTATION TECHNIQUES

The methods used for authorship analysis include univariate, multivariate statistics. The first method of computation was a univariate approach, which failure leads to the advent of a multivariate approach and the new machine learning algorithms proposed by Koppel [33]. Machine-learning algorithms are statistics of multivariate nature. There are some examples of machine learning algorithms; such as Support Vector Machine (SVM), Neural Network (NN), and decision trees proposed by Iqbal [29].

6.1 Univariate Methods

Scientific analysis of authorship can be dated to the late nineteenth century, and the main idea was that authorship could be determined by the relation of word length to the relative frequency of occurrence. Some statistical researchers attempted to identify fixed properties in written texts in the early twentieth century, which gave rise to an idea of evaluating authorship that these fixed features could be used to solve authorship problems, whereas static features were used later to evaluate authorship. Being ineffectual and giving way to a multivariate approach proposed by Koppel [33].

6.2 Multivariate Methods

In 1964, Mosteller [13] mentioned new ways to solve the problem of authorship attribution by combining multiple stylometric features, and it was believed that this was the beginning of the multivariate approach. Mosteller and Wallace are said to have mainly used functional words that were content-independent and applied Bayesian classification techniques to solve the problem of author attribution, and the result was somewhat reliable. The development of machine learning techniques, in particular text classification techniques, enabled an investigation of authorship [33]. The aim of using text categorization techniques to solve the question of authorship analysis is to turn the training dataset into feature vectors and to use text categorization techniques to set target class boundaries and outliers.

In multi-dimensional space; texts can be seen as vectors. Statistical and machine learning methods such as Discriminant Analysis, Support Vector Machines, Decision Trees, Neural Networks, Genetic Algorithms, Memory-based learners, Ensemble Classification Methods can be used to train classification models [9].

7. PERFORMANCE MEASURES

The confusion matrix shown in Table 1. The confusion matrix is the source of the classification problem performance measures [49]. Where True Positive (TP) is predicted values correctly predicted as actual positive, False Positive (FP) is predicted values incorrectly predicted an actual positive. i.e., negative values predicted as positive, False Negative (FN) is a positive value predicted as negative, and True Negative (TN) is a predicted value correctly predicted as an actual negative. The performance metrics which allows us to measure Accuracy, Precision, True Positive Rate (RECALL or Sensitivity), F1-measure (F1 Score), G-mean, and the True Negative Rate (Specificity). The formulations for each one are shown in Table 2.

Table 1: Confusion Matrix of the Performance Measures

		True Class	
		True	**False**
Predict Class	**Positive**	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

- **True Positive Rate TPR (Sensitivity)**

It is called also RECALL (REC). It represents the degree of obtaining relevant information. Thus, by dividing True Positive Count (relevant information obtained) by some of True Positive Count and False Positive Count (irrelevant information obtained), the precision is determined. Calculated as the number of true positive predictions (TP) divided by the total number of positive observed (TP+FN).

- **True Negative Rate (Specificity)**

It is also called True Negative Rate (TNR), it calculated as the number of true negative predictions (TN) divided by the total number of negative observed (TN+FP).

- **Precision**

It is called also a Positive Predictive Value (PPV), it used in information retrieval and other pertinent fields. It is suggested to what degree the model can collect more relevant information than non-relevant information. Precision calculated as the number of true positive predictions (TP) divided by the total number of positive predictions (TP + FP).

- **Accuracy**

The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value. Calculated as the ratio of the correct decisions (TP + TN) divided by the total population (TP + FP + TN + FN).

- **F1-measure**

It is also called F1- Score. It is the harmonic mean or sub-contrary mean of precision and Recall. It is used to evaluate a model by balancing precision and recall.

- **G-mean**

It is the geometric mean of recall and precision; it is a performance metric that calculates a square of multiplication of true positive rate and positive predictive value.

Table 2: Performance Measures

Performance Measure	Formula Equation
True Positive Rate (RECALL or Sensitivity)	$\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$
True Negative Rate (Specificity)	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1-measure	$2x\frac{(\text{Recall } x \text{ Precision})}{(\text{Recall } + \text{ Precision})}$
G-mean	$\sqrt{\text{Recall } x \text{ Precision}}$

8. EVALUATION

Normally one author only has one book in the book collection corpus. Thus, the variety of content of different books in different fields may contribute to the differentiation of different authors rather than the writing styles of the authors. To solve the real-life problems of forensic authorship verification these models are not very likely to be applied, because the length of words still too long and, unsatisfied for applying in short text like Twitter. As regards usability, models of one-class classification are more likely to be applied to solve specific problems. The need to gather outlier data makes language-dependent on the two-class classification models. Moreover, outlier data should be as close as possible to the target class, which is another limitation. However, the one-class classification models do not require an outlier class representation.

9. CONCLUSION

We introduced authorship verification which used some approaches to obtain a good performance. Both the approaches of one-class classification and two-class classification were used to solve the problem of verification. In addition, both the profile-based approach and an instance-based approach were introduced to solve the problem, whereas the instance-based approach is becoming more prevalent. In terms of computational techniques, various forms of machine learning techniques were introduced by researchers.

In the future there several important questions are still open for the authorship verification issue. Perhaps the most crucial issue is the text length. Stylometric features are not yet able to represent adequately the stylistic choices of texts. Hence, they can be used only as a complement in other, more powerful features coming from the lexical or the character level.

10. REFERENCES

[1] Da Silva, N.F., Hruschka, E.R., and Hruschka Jr, E.R. 2014 Tweet sentiment analysis with classifier ensembles. Decision Support Systems.

[2] Juola, P., Jr., J.I.N., Stolerman, A., Ryan, M.V., Brennan, P., and Greenstadt, R. 2013 A Dataset for Active

Linguistic Authentication, Book A Dataset for Active Linguistic Authentication.

[3] Halteren, H.v. 2004 Linguistic profiling for authorship recognition and verification, Book Linguistic profiling for authorship recognition and verification.

[4] Koppel, M., and Schler, J. 2004 Authorship verification as a one-class classification problem. Authorship verification as a one-class classification problem.

[5] Stein, B., Lipka, N., and zu Eissen, S.M. 2008 Meta analysis within authorship verification, Book Meta analysis within authorship verification.

[6] Kestemont, M., Luyckx, K., Daelemans, W., and Crombez, T. 2012 Cross-genre authorship verification using unmasking. English Studies.

[7] Mosteller, F., and Wallace, D.L. 1963 Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association.

[8] Carbonell, J.G., Michalski, R.S., and Mitchell, T.M. 1983 An overview of machine learning, Machine learning (Elsevier).

[9] Stamatatos, E. 2009 A survey of modern authorship attribution methods, Journal of the American Society for information Science and Technology.

[10] Potha, N., and Stamatatos, E. 2014 A profile-based method for authorship verification, Book A profile-based method for authorship verification (Springer).

[11] Kešelj, V., Peng, F., Cercone, N., and Thomas, C. 2003 N-gram-based author profiles for authorship attribution, Book N-gram-based author profiles for authorship attribution.

[12] Ma, J., Xue, B., and Zhang, M. 2016 A profile-based authorship attribution approach to forensic identification in Chinese online messages, Book A profile-based authorship attribution approach to forensic identification in Chinese online messages (Springer).

[13] Mosteller, F., and Wallace, D. 1964 Inference and Disputed Authorship The Federalist Reading, Massachusetts: Addison-Wesley.

[14] Argamon, S., Šaric, M., and Stein, S.S. 2003 Style mining of electronic messages for multiple authorship discrimination: first results, Book Style mining of electronic messages for multiple authorship discrimination: first results.

[15] Koppel, M., and Schler, J. 2003 Exploiting stylistic idiosyncrasies for authorship attribution, Book Exploiting stylistic idiosyncrasies for authorship attribution.

[16] De Vel, O., Anderson, A., Corney, M., and Mohay, G. 2001 Mining e-mail content for author identification forensics, ACM Sigmod Record.

[17] Zheng, R., Li, J., Chen, H., and Huang, Z. 2006 A framework for authorship identification of online messages: Writing-style features and classification techniques, Journal of the American society for information science and technology.

[18] Jankowska, M., Milios, E., and Keselj, V. 2014 Author verification using common n-gram profiles of text documents, Book Author verification using common n-gram profiles of text documents.

[19] Li, Z. 2013 An Exploratory Study on Authorship Verification Models for Forensic Purpose.

[20] Brocardo, M.L., Traore, I., Saad, S., and Woungang, I. 2013 Authorship verification for short messages using stylometry, Book Authorship verification for short messages using stylometry (IEEE).

[21] Halvani, O., Steinebach, M., and Zimmermann, R. 2013 Authorship verification via k-nearest neighbor estimation, Notebook PAN at CLEF.

[22] Koppel, M., Schler, J., and Bonchek-Dokow, E. 2007 Measuring differentiability: Unmasking pseudonymous authors, Journal of Machine Learning Research.

[23] Nirk, S. 2016 ship Veri', Procedia Computer Science.

[24] Boenninghoff, B., Hessler, S., Kolossa, D., and Nickel, R.M. 2019 Explainable authorship verification in social media via attention-based similarity learning, Book Explainable authorship verification in social media via attention-based similarity learning (IEEE).

[25] Lambers, M., and Veenman, C.J. 2009 Forensic authorship attribution using compression distances to prototypes, Book Forensic authorship attribution using compression distances to prototypes (Springer.).

[26] Loper, E., and Bird, S. 2002 NLTK: the natural language toolkit.

[27] Chen, X., Hao, P., Chandramouli, R., and Subbalakshmi, K. 2011 Authorship similarity detection from email messages, Book Authorship similarity detection from email messages (Springer).

[28] Vilariño, D., Pinto, D., Gómez, H., León, S., and Castillo, E. 2013 Lexical-syntactic and graph-based features for authorship verification, Book Lexical-syntactic and graph-based features for authorship verification.

[29] Iqbal, F., Khan, L.A., Fung, B.C., and Debbabi, M. 2010 E-mail authorship verification for forensic investigation, Book E-mail authorship verification for forensic investigation.

[30] Hoover, D.L. 2003 Another perspective on vocabulary richness, Computers and the Humanities.

[31] Stamatatos, E. 2017 Authorship attribution using text distortion, Book Authorship attribution using text distortion.

[32] Grieve, J. 2007 Quantitative authorship attribution: An evaluation of techniques, Literary and linguistic computing.

[33] Koppel, M., Schler, J., and Argamon, S. 2009 Computational methods in authorship attribution, Journal of the American Society for information Science and Technology.

[34] Tanguy, L., Urieli, A., Calderone, B., Hathout, N., and Sajous, F. 2011 A multitude of linguistically-rich features for authorship attribution, Book A multitude of linguistically-rich features for authorship attribution.

[35] Miller, G.A. 1998 WordNet: An electronic lexical database, (MIT press).

[36] Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., and Levitan, S. 2007 Stylistic text classification using functional lexical features, Journal of the American Society for Information Science and Technology.

[37] Halliday, M.A.K., and Matthiessen, C.M. 2013 Halliday's introduction to functional grammar, (Routledge).

[38] McCarthy, D. 2006 Relating WordNet senses for word sense disambiguation, Book Relating WordNet senses for word sense disambiguation.

[39] Zhang, D., and Lee, W.S. 2006 Extracting key-substring-group features for text classification, Book Extracting key-substring-group features for text classification.

[40] Marton, Y., Wu, N., and Hellerstein, L. 2005 On compression-based text classification, Book On compression-based text classification, (Springer)

[41] Keogh, E., Lonardi, S., and Ratanamahatana, C.A. 2004 Towards parameter-free data mining, Book Towards parameter-free data mining.

[42] Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P.M. 2004 The similarity metric, IEEE transactions on Information Theory.

[43] Cilibrasi, R., and Vitányi, P.M. 2005 Clustering by compression, IEEE Transactions on Information theory.

[44] Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. 2001 An information-based sequence distance and its application to whole mitochondrial genome phylogeny, Bioinformatics.

[45] Chen, X., Francia, B., Li, M., Mckinnon, B., and Seker, A. 2004 Shared information and program plagiarism detection, IEEE Transactions on Information Theory.

[46] Sculley, D., and Brodley, C.E. 2006 Compression and machine learning: A new perspective on feature space vectors, Book Compression and machine learning: A new perspective on feature space vectors, (IEEE).

[47] Forman, G. 2003 An extensive empirical study of feature selection metrics for text classification, Journal of machine learning research.

[48] Sebastiani, F. 2002 Machine learning in automated text categorization, ACM computing surveys (CSUR).

[49] Olson, D.L., and Delen, D. 2008 Advanced data mining techniques, Springer Science & Business Media.