Author Verification Using Rich Set of Linguistic Features

A. Bhanu Prasad, S. Rajeswari, A. Venkannababu and T. Raghunadha Reddy

Abstract Author Verification is a type of author identification task, which deals with identification of whether two documents were written by the same author or not. Mainly, the detection performance depends on the used feature set for clustering the documents. Linguistic features have been utilized for author identification according to the writing style of a particular author. Disclosing the shallow changes of the author's writing style is the major problem which should be addressed in the domain of authorship verification. It motivates the computer science researchers to do research on authorship verification in the field of computer forensics. In this work, three types of linguistic features such as stylistic, syntactic, and semantic features are used to improve the accuracy of author verification. The Naïve Bayes multinomial classifier is used to build the classification model and good accuracy is achieved for Author Verification.

Keywords Author verification · Stylistic features · Syntactic features
Semantic features · Naïve Bayes multinomial

A. Bhanu Prasad (✉)
CSE Department, Vardhaman College of Engineering, Hyderabad, India
e-mail: andrajub4u@gmail.com

S. Rajeswari
CSE Department, VR Siddhartha Engineering College, Vijayawada, India
e-mail: rajeswari.setti@gmail.com

A. Venkannababu
CSE Department, Sri Vasavi Engineering College, Tadepalligudem, AP, India
e-mail: venkannababu.alamuru@gmail.com

T. Raghunadha Reddy
IT Department, Vardhaman College of Engineering, Hyderabad, India
e-mail: raghu.sas@gmail.com

S. C. Satapathy et al. (eds.), *Information and Decision Sciences*,
Advances in Intelligent Systems and Computing 701,
https://doi.org/10.1007/978-981-10-7563-6_21

1 Introduction

There is a vast amount of data on the Internet and it is growing rapidly every day. Such a high rate of growth also brings some problems with it. Fraudulent, stolen, or unidentified data are encountered online on a daily basis. These problems can be dangerous and serious problems in places like the public websites, government, forensics, and schools. Because of these threats, and in detection of truth, it is important to know the author of a text.

Authorship Analysis is divided into three categories including Authorship Attribution, Authorship Verification, and Authorship Profiling. Authorship Attribution studies a text in dispute and finds the corresponding author in a set of candidate authors. Authorship Verification compares multiple pieces of written text and determines whether they are written by the same author or not without identifying the author. Authorship Profiling detects unique characteristics like gender, age, location, nativity language, and educational background of an author's written texts and creates an author profile. In this work, the Author Verification task is concentrated. Author Verification techniques are important in several information processing applications.

In the context of cyberspace, a digital document found can be used as an evidence to prove that a suspect is a criminal if he/she is the author of the document. If the suspect authors are unknown, i.e., there is no suspect, thus this is commonly known as an authorship identification problem. However, there are also some cases when the identification of the author is not necessary, i.e., it is enough just to know if the document in dispute was written by the author of the documents that are given. This is a problem faced by many forensic linguistic experts which are called as authorship verification problem.

This paper is organized as follows. Section 2 demonstrates the existing approaches already implemented and tested in authorship verification. Section 3 introduces the set of linguistic features used for document representation in authorship verification. The classification procedure and our approach for finding accuracy of author verification are explained in Sect. 4. In Sect. 5, the experimental results obtained will be discussed and Sect. 6 presents the conclusion.

2 Literature Review

Authorship Verification is the process of verifying an author by checking whether the document is written by the suspected author or not [1]. Victoria Bobicev proposed [2] a method to automatically detecting the author of a given text when the corpus contains small training sets with known authors. They used the prediction by partial matching (PPM) method based on statistical n-gram model. Without feature engineering, PPM obtains total information from the original corpus. They experimented with a corpus of 30 authors, 100 posts of each author and approximately each post length is 150–200 words. It was observed that their system accuracy measure F-measure is not increased when the document length was increased.

Vanessa Wei Feng et al. adopted [3] an unmasking approach, which is used to enhance the quality of features used in building weak classifiers. They experimented with 538 features for English, 568 for Greek, and 399 for Spanish language. The features include coherence features and stylometric features. They observed that their work achieved best accuracy for English and Spanish texts, but less accuracy for Greek texts.

Darnes Vilariño et al. used [4] syntactic, lexical, and graph based features to represent the document vectors. Subdue data mining tool is used to extract the graph-based features. A support vector machine is used to prepare the classification model. Lexical-syntactic features include phrase level features such as word suffixes, stopwords, punctuation marks and trigrams of POS, and character level features such as vowel combination and vowel permutation. It was observed that their system run time is greater than most of the other submissions.

Cor J. Veenman et al. used [5] the compression dissimilarity measure to compute the compression distance between the documents. They proposed three approaches such as nearest neighbor with compression distances, two class classifications in compression prototype space and bootstrapped document samples for author verification task. It was observed that they obtained best accuracy among the submissions in PAN 2013 competition.

Michiel van Dam used [6] the profile-based approach and they applied common N-gram (CNG) method which utilized the normalized distance measure between short and unbalance text. In CNG method, each document is represented with character n-grams. It was observed that their approach obtained good accuracy for English and Spanish languages, but fails for Greek language.

Shachar Seidman proposed [7] a general impostors method which is based on comparing the similarity between given documents and number of external documents. It was observed that their approach achieved overall first rank in the competition. Timo Petmanson extracted [8] frequent significant features such as nouns, punctuations, verbs, and first words of sentences or lines, they used principal component analysis to compute the Matthews correlation coefficient for all pairs of extracted features.

Alberto Bartoli et al. proposed [9] a machine learning approach by using a set of linguistic features. They extracted various features such as word n-grams, character n-grams, POS tag n-grams, word lengths, sentence lengths, sentence lengths n-grams, word richness features, punctuation n-grams, and text shape n-grams. Their approach obtained first rank in author verification for Spanish language in PAN 2015 competition.

3 Linguistic Features

A feature is an attribute of an object that can characterize the document. Most objects and entities have more than one feature. In machine learning, such objects are represented as a vector of features. Features help us to differentiate the objects from one another and also help to describe them. It is essential to select useful and distinctive features in order to achieve high classification scores. In this work, the experimentation carried out with numeric and semantic features and also experiment on each type of feature in isolation as well as experimenting by merging them together gradually.

A numeric feature is a measurement. Numeric features represent a feature of a document with numbers. Two types of numerical features such as stylistic features and syntactic features are used in our experiment. For example, the word count in a document is a numeric feature which contains numeric values. The following set of numerical features was used for the experiments covering almost all the aspects of the previously defined stylistic features in the literature. Typically, these stylistic features include total number of characters, average length per word, number of sentences, words per sentences, words longer than six characters, total number of short words, number of syllables, syllables per word, number of complex words (more than 3 syllables), number of capital letters, number of small letters, ratio of capital letters to small letters, capital letters words, number of words, contraction words, the number of words with hyphens, words followed by digits, unique terms, ratio of number of words which contain more than 3 syllables to total number of words, number of acronyms, number of foreign words, number of words that occur twice (hapax dis legomena), and number of specific words.

Syntactic features include part of speech based features such as number of nouns, number of passive verbs, number of base verbs, number of adjectives, number of clauses and number of phrases, number of articles, number of prepositions, number of coordinate conjunctions, and number of auxiliary verbs. In this work, another syntactic measure such as punctuation measures which is not in the literature as important and those includes number of commas, number of colons (:), number of semicolons (;), number of single quotes (‘), number of double quotes (“), number of exclamation marks (!), number of question marks (?) and the number of “etc.”. Syntactic features have been extracted by using the parse trees of the sentences. These parse trees are obtained by using the Stanford Parser.

A numeric feature was representing features with numbers. Semantic features represent features with sets of meanings. Synonym sets are used as semantic features. The semantic features are used to directly tie the features to the meaning of word. The meaning of the words is used as semantic features. A WORDNET of synonym set is created for each author as a model which can represent an author’s writing topic. To use synonyms for semantic features, WORDNET is needed.

For our work, the experimentation is carried out on PAN 2014 competition author verification dataset. Table 1 shows the characteristics of the corpus used in our work.

Table 1 Dataset characteristics of PAN 2014 competition for English language

Features	Testing data	Training data
Number of authors	100	100
Number of documents	100	500
Vocabulary size	12764	41583
Number of documents per author	1	5
Average words per sentence	21	25
Average words per document	1121	1135

4 Our Approach

The procedure of author verification process is represented in Fig. 1. In this procedure, first, the preprocessing techniques such as stopwords removal and stemming are performed on the collected corpus. Then, the features that differentiate the writing style of the author from the updated corpus are extracted. The document vectors are generated by using extracted features from the corpus. The document vectors are given to classification algorithm to generate the classification model. Finally, the classification model is used to analyze the unknown document and predicts whether the document is written by the particular author or not.

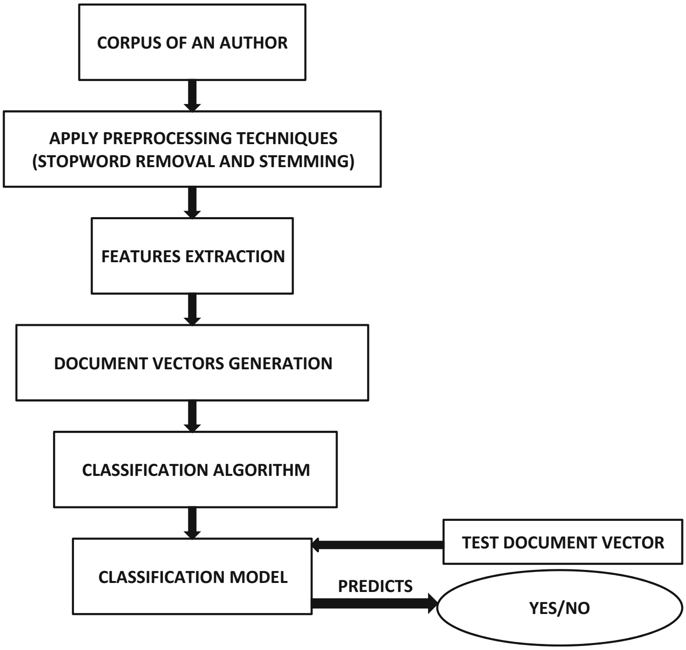

Fig. 1 The procedure of our approach for author verification

Classification is a problem of identifying which category a new input belongs in. An algorithm that implements a classification is called a classifier. There are many classification algorithms like Naïve Bayes multinomial, random forest, decision trees, bagging, support vector machines, and neural networks used for classification. In this work, Naïve Bayes multinomial classifier is used to predict the accuracy of author verification.

K-fold cross-validation testing method is used to test our classifier. K-fold cross-validation is widely used for classifiers. K-fold cross-validation has K iterations. On each iteration, one random unit is selected for testing and the remaining K-1 is used for training. This process is repeated K times while each randomly selected unit is used exactly once. With this method, we ensure that all data is used for both training and testing.

5 Empirical Evaluations

In this work, experimentation carried out with machine learning practices to solves this problem. Naïve Bayes multinomial classifier is identified to generate a efficient classification model because it has high scalability due to number of features/predictors it can have. At the beginning of each classification, the documents are transformed into feature vectors. This transformation/extraction process is performed only once. After extracting the features from the documents, add these features to a feature vector. As discussed above, in this classification process, 10-fold cross-validation is used. For each fold of the validation, construct the training vectors and test vectors. The training vectors are passed to the classifier and the classifier will create a classification model by iterating all the training vectors. This classification model is used to test the test vector and calculate the efficiency of classifier. The average of all folds will be the final accuracy of our classification methodology.

Precision and recall measures are used as evaluation measures to find the accuracy of our approach. Precision is the ratio of number of problems that correctly answered to total number of problems. Recall is the ratio of number of problems that correctly answered to total number of answers. The accuracies of various combinations of features are represented in Table 2. The combination of stylistic features, syntactic, and semantic features achieved the good precision of 97.8% and recall of 96.7% by using Naïve Bayes multinomial classifier.

Table 2 The accuracies of precision and recall measures for various set of features

Features used	Precision (%)	Recall (%)
Stylistic features	86.8	84.1
Syntactic and POS features	88.2	90.7
Semantic features	91.6	92.9
Combination of all features	97.8	96.7

6 Conclusion

The selections of features vary by the nature of the document. In this work, experimentation carried out with three types of features such as stylistic features, syntactic features, and semantic features. Our work obtained good precision of 97.8% for Author Verification by using Naïve Bayes multinomial classifier.

References

1. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inform. Sci. Technol. **60**(1), 9–26 (2009)
2. Bobicev, V.: Authorship detection with PPM. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
3. Feng, V.W., Hirst, G.: Authorship verification with entity coherence and other rich linguistic features. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
4. Vilariño, D., Pinto, D., Gómez, H., León, S., Castillo, E.: Lexical-syntactic and Graph-based features for authorship verification. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
5. Veenman, C.J., Li, Z.: Authorship verification with compression features. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
6. van Dam, M.: A basic character n-gram approach to authorship verification. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
7. Seidman, S.: Authorship verification using the impostors method. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
8. Petmanson, T.: Authorship identification using correlations of frequent features. In: Proceedings of CLEF 2013 Evaluation Labs (2013)
9. Bartoli, A., Dagri, A., Lorenzo, A.D., Medvet, E., Tarlao, F.: An author verification approach based on differential features. In: Proceedings of CLEF 2013 Evaluation Labs (2015)