ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Maximum Likelihood Text Classification Algorithm Using Machine Learning For Authorship Attribution

Mercy D'cruz A

M. Phil, Dept. of Computer Science, Indian Institute of Information Technology and Management, Kerala, India

ABSTRACT: Authorship attribution is a powerful technique and can be useful for forensic scientists. In this paper wereport a detailed investigation in the field of authorship attribution. The focus of this work is to propose Maximum likelihoodalgorithm in the context of short texts. We analyzed basic-9 and writeprints(limited) feature sets, using a tool and corpusalready available. Experiments on a number of short texts using writeprints(limited) feature set given promising results. According to F-measure writeprints(limited) provide better generalization performance than basic-9.

KEYWORDS: Authorship attribution; Maximum likelihood algorithm; classification; stylometry; basic-;writeprints(limited).

I. INTRODUCTION

Linguists Peter Millican and Patrick Juola uncovered thatJ.k. Rowling was the author of the novel The Cuckoos calling, which was published under the nick name Robert Galbraith in year 2013. The UNiversity & Airline BOMber (UNABOM)was recognized as Ted Kaczynski from his anonymously published document Unabomber Manifesto by analysing thewriting style. Writing analysis area is currently dominated byAI such as pattern recognition and neural networks. Thesedevelopments in AI made authorship analysis a necessarynew research area in security [2], [14]. Authorship analysisis the technique of inferring the authorship of a documentby analyzing the writing styles from the document content.In linguistic field, authorship analysis has its root known asstylometry. Authorship analysis studies can be classified intocategories such as:

- Authorship attribution is the process that identifies the Maximum likelihood of a particular author having written a piece of text by examining other sample texts produced by that author.
- Authorship profiling is the process that identifies the characteristics of the author that produced a given piece of text. These characteristics include gender, education,cultural background and language familiarity.
- Similarity detection compares multiple pieces of text samples and determines whether or not they are produced by a single author. Similarity is commonly used in the context of plagiarism detection which involves the partial or complete duplication of a piece of text with or without permission of the original author [1],[13], [14].

II. RELATED WORK

Globally, significant progress has been achieved in authorship attribution area. The challenge in this field is to identify an author, when the text sample is short or the number of authors increases. Chaski (2005) achieved 95.70% accuracy on authorship attribution with 10 authors as the evaluation sample [12]. Iqbal et al. (2008) proposed AuthorMiner approach which achieved an accuracy of 80.5% with 6 authors and 77% with 10 authors [10]. Hadjidj et al. (2009) obtained classification rates 77% and 71% for sender identification,73% and 69% for sender recipient identification, and 83% and 83% for sender-cluster identification using C4.5 and SVM classifers to determine authorship [6]. Iqbal et

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

al. (2010) also reported 90% accuracy with 3 authors using k-means for author attribution. Reported accuracy decreased to 80% with the increase in number of authors as 10 [4].
The main goal of this paper is to propose a Maximum likelihood algorthim for authorship attribution which has not mentioned as an algorithm in this field. The paper is organized as follows : Section III briefly outline the Maximum Likelihood Algorthim. Section IV presents experimental methodology and feature sets used in the experiments. Section V reviews experimental results and discussion. Finally, Section VI concludes withgeneral experiments and future scope for the work.

III. PROPOSED ALGORITHM

The proposed algorthim is based on the concept of likelihood function. The likelihood function quantifies the uncertainty of the parameters taking particular values and produces observed realizations.

Algorithm :*Maximum Likelihood Algorithm*

Input : Training document set for each known author and testing document set.
Output : Maximum Likelihood author is identified for the unknown document which is assumed to be written by one of the author from training set.

Step 1 : Consider A= {A_1,A_2, ... A_N} where N be the no: of authors. For each author, let Tr be the training document set of short texts where r =1, 2,... n. Let Et be the test document set of short texts where t = 1,2,.... n.

Step 2 : **If**the presented training document and test documentis not same (Tr != Et).

Do the following steps:
2.1: **[Pre-processing]** Data pre-processing is done for each training document Tr
and the test document Et.
2.2: **[Feature]** Let F be the feature set. Select a feature f1 where f1 2 F.
2.3: **[Classifier]** Let C be the classifier method set. Select a classifier c1 where c1 2 C.
2.4: **[Classification]** Classify data using selected feature f1 and classification method c1. Using the feature vector the Maximum likelihood author of the test document is identified and displayed as output.
2.5: **[Add classifier]** To add more classifier goto step 2.3.
2.6: **[Add feature]** To add more features goto step 2.2.
2.7: Repeat until the training document and test document is classified correctly to identify author.

Step 3 : **Else**
Display selected training document and test documentis same. Goto step 2.
Step 4 : **Exit**

IV. METHODOLOGY

Our writing style anonymization framework uses JStylo. JStylo is a independent authorship attribution platform. The NLP techniques are used to extract features from text samples. The extracted features of texts are classified using machine learning methods [3] [14]. JStylo initially grasp the style of known candidate authors based on texts of those authors, then features authorship of the unknown texts to any of the knownauthors. The work-flow consists of four stages consists of defining a problem set, feature selection, classifiers selection and running the analysis. A feature set is defined by a set of various stylistic features to be extracted from the text samples.JStylo supports pre-defined feature sets such as Basic-9 TableI and WritePrints(Limited) Table II [3], [5], [8], [11].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

TABLE I. BASIC 9 FEATURE SET

Category
Unique Words Count
Complexity
Sentence Count
Average Sentence Length
Average Syllables in Word
Gunning-Fog Readability Index
Character Space
Letter Space
Flesch Reading Ease Score

TABLE II. WRITEPRINTS FEATURE SET

Category	Description
Character related	Total characters, percentage of letters, percentage of digits, percentage of uppercase letters, etc. and frequency of character unigram, most common bi-grams and tri-grams
Digits, punctuations, special characters	Frequency of digits (0-9), punctuations and special characters (e.g., %,&, *)
Word related	Total words, number of characters per word, frequency of large words, etc. Most frequent word uni-/bi-/ tri-grams
Function words and parts-of-speech	Frequency of function words and parts -of-speech

The Basic-9 feature setTable I contains nine features that were used for experiments and Writeprints(Limited) feature set Table II contains features used for the Writeprints technique [3], [5], [8], [11]. The training documents set are mined for the selected features, which are later used for training the classifier. The same features are mined in the test set, for later classification by the trainedclassifiers. For each feature, text pre-processing(optional) thatallow various methods such as stripping all punctuation which is to be applied before the feature extraction. The gist of the feature which is the feature extractor itself and the feature post-processing(optional) which is to be applied on the features such as picking the top features frequency-wise after extraction. There are various analysisconfigurations available in Jstylo. The major choice is to run a 10-fold cross validationanalysis over the training corpus or to train the classifiers using a training corpus and classifying the test documents. The classifiers available for selection are a subset of Weka classifiers commonly used, such as support vector machine smo etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

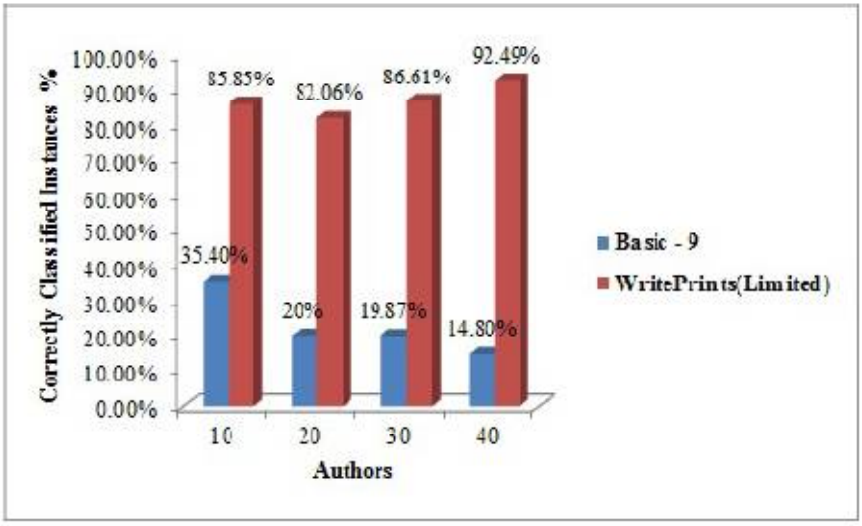
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

V. RESULTS AND DISCUSSION

Reuter 50 50 dataset is used for experiments [16]. It contains 50 authors and 50 documents per author. The 40 authors are selected to form the corpus. The training corpus contains 2,000 texts and test corpus also consists of 2000 text which is non-overlapping with training texts. The experiments with JStylo were conducted using a SVM classifier, over two feature sets - the Basic-9 feature set and the WritePrints (Limited). The corpus is evaluated using 10-folds cross-validation [7], [9]. The 10-fold cross-validation is the process of randomly partitioning the training data into 10 equal-sized parts. The classification is performed 10 times, each time a different part is used as test data while the remaining 9 parts are used as training data. Each part is used as test data exactly once. The classification task results are averaged. Therefore, this method reduces the instability of the classification. The results of 10-fold crossvalidation are summarized in Fig 1.

Fig.1.10-Folds Cross-Validation



The Weka output includes several measures that indicate the performance of the classification. These measures can also be used to compare the basic feature set and writeprint using the SVM SMO classifier. The effective measure is the percentage correctly classified instances. The correctly classified instances is used as the valuable measure to compare the best feature set for the classification task. The Weka outputalso includes a confusion matrix to evaluate accuracy.

Although the Basic-9 feature set did not produce as high results as compared Writeprints (Limited) feature set. The performance is evaluated by fixed number of test documents (50 per author) and varying the number of authors from 10 to 40 with a difference of 10.

TABLE III. PERFORMANCE MEASURES FOR SVM SMO CLASSIFIER WITH BASIC-9 FEATURE SET

Measures	10	20	30	40
Accuracy	0.432	0.319	0.252	0.306
Precision	0.434	0.322	0.253	0.308
Recall	0.432	0.319	0.252	0.306
F-measure	0.432	0.320	0.252	0.306

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

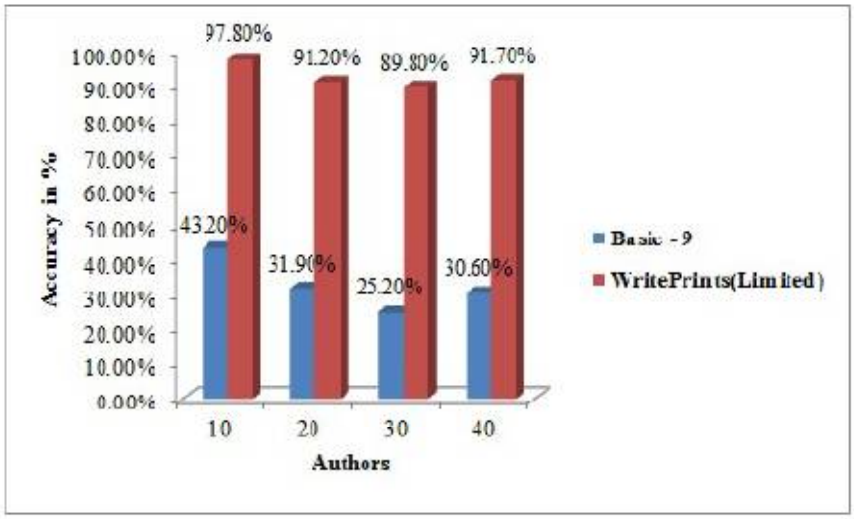
Vol. 4, Issue 11, November 2016

TABLE IV. PERFORMANCE MEASURES FOR SVM SMO CLASSIFIER WITH WRITEPRINTS FEATURE SET

Measures	10	20	30	40
Accuracy	0.978	0.912	0.898	0.917
Precision	0.978	0.913	0.90	0.921
Recall	0.976	0.912	0.898	0.917
F-measure	0.97.9	0.912	0.898	0.918

It is notable that using the Writeprints (Limited) feature set given promising results across all experiments. The confusion matrix are generated to evaluatemeasures such as accuracy, precision, recall, F-mearsure [15]. The performance evaluation are shown as in Table III and Table IV. The accuracy in percentage are plotted in Fig. 2.

Fig.2.Accuracy obtained with fixed the number of test documents as 50(per author) and varying the number of authors from 10 to 40 with a difference of 10



Accuracy is used to indicate the number of correctly classifiedinstances over the total number of test instances by calculating the average of accuracy, as in Eq.(1).

$$\text{Accuracy} = \frac{\text{Number of documents that are correctly classified}}{\text{Total Number of documents}} \quad (1)$$

Comparsion with the existing works of chaski (2005), Iqbal et al. (2008), Iqbal et al. (2010), our work obtained an accuracy of 97.8% for 10 and 91.7% for 40 authors [12], [10], [4].

VI. CONCLUSION AND FUTURE SCOPE

In this work, Maximum likelihood Algorithm for authorship attribution is introduced based on the likelihood concept. The SVM classifier is used. Experiments of 10-folds crossvalidation have been done separately on a Reuter 50 50dataset using an SVM SMO classifier. Basic-9 features and writeprints(Limited) features are extracted from this dataset. The writeprints(limited) features are better than the Basic-9 features, depending on the average accuracy of the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

experiments done. The Writeprints(limited) gave the best score for this classifier obtained up to 92.49% classification accuracy in 10-fold cross validation. By analyzing the confusion matrix, observed 91.70% accuracy for 40authors ofshort texts. This performance can be further improved by adding more features and using different classifiers.

REFERENCES

1. M. L. Brocardo; I. Traore; S. Saad; and I. woungang, "AuthorshipVerification for Short Messages Using Stylometry", In Proc. Of the IEEE Intl. Conference on Computer, Information and TelecommunicationSystems (CITS 2013), Piraeus-Athens, Greece, May 7-8, 2013(Best PaperAward).
2. A. Abbasi and H. Chen, "Applying authorship analysis to extremist group web forum messages", IEEE Intelligent Systems, 20:67-75, 2005.
3. M. Koppel and J. Schler, "Authorship verification as a one-class classification problem, In Proc. Of the 21st international conference on Machine learning, ICML '04, page 62, 2004.
4. F. Iqbal; H. Binsalleeh; B. C. Fung and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation, Digital Investigation, 7(1-2):56 - 64, 2010.
5. O. Canales; V. Monaco; T. Murphy; E. Zych; J. Stewart; C.T. A. Castro; O. Sotoype; L. Torres; and G. Truley, "A stylometry system for authenticating students taking online tests", CSIS, Pace University, 2011.
6. R. Hadjidj; M. Debbabi; H. Lounis; F. Iqbal; A. Szporer; and D. Benredjem, "Towards an integrated e-mail forensic analysis framework", Digital Investigation, 5(3-4): 124 - 137, 2009.
7. F. Iqbal; H. Binsalleeh; B. C. Fung; and M. Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications", Information Sciences,231:(98 -112), 2011.
8. A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identify level identification and similarity detection in cyberspace", ACM Transactions on Information Systems, 26:2, 2008.
9. Krsul and E. H. Spafford, "Authorship analysis: Identifying the author of a program", Computers and Security, 16(3):233 - 257, 1997.
10. F. Iqbal; R. Hadjidj; B. C. Fung; and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics". In Proc. Of the Eighth Annual DFRWS Conference, Digital Investigation, 5, Supplement(0):S42 S51, 2008.
11. M. Koppel; J. Schler; and S. Argamon, "Authorship attribution in the wild", Language Resources and Evaluation, 45(1): 83-94, 2011.
12. Chaski and E Carole, "Whos at the keyboard? Authorship attribution in digital evidence investigations".International Journal of Digital Evidence, 4(1):1-13, 2005.
13. E. Stamatatos, A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3):538556, 2009.
14. P. Juola. Authorship attribution, Foundations and Trends in information Retrieval,1(3):233334, 2006.
15. D. V. Olivier, "Mining e-mail authorship". In Proc. Of the workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining, 2000.
16. Reuter 50 50 Dataset, UCI Machine Learning Repositiory, https://archive.ics.uci.edu/ml/datasets/Reuter 50 50.

BIOGRAPHY

Mercy D'cruz A received Mphil in Computer Science , Indian Institute of Information Technology and Management - Kerala, Affiliated by Cochin University of Science and Technology, kerala, India. She received Master of Computer Application (MCA) degree from Cochin University of Science and Technology,kerala, India. Her research interests are Artifical Intelligence , Authorship Analysis, Machine Learning etc.