ISSN:0975-9646

Stylometric Approach For Author Identification of Online Messages

Ms.Smita Nirkhi
Research Scholar
G.H.Raisoni College of Engineering
Nagpur, India

Dr.R.V.Dharaskar
Director
MPGI
Nanded, India

Dr.V.M.Thakre
Professor & Head
Government Engineering,
Amravati

Abstract— **In the era of internet, the use of online blogs, forum, social network and email is very popular for communication. At the same time due to anonymity, cybercriminals making use of these online messages for illegal activities like cyber bulling, fishing etc.In this context, Authorship identification plays important role by finding the plausible author of anonymous text .Authorship Identification thus helps to recognize author of unknown text. This paper will focus on the use of Stylometry approach along with n-gram as feature for finding unique write prints of authors. The main steps involved for this experimentation are feature extraction, classification and author identification.SVM classifier is used to check accuracy of this approach.**

Keywords—SVM, Stylometry, N-Gram, Z-score

I. Introduction

Stylometry is the study of the unique linguistic styles and writing behaviors of individuals in order to determine authorship [1]. The main hypothesis principal of stylometric studies is that authors have an unconscious as well as a conscious aspect to their writing style. Every author's style is thought to have certain features that are independent of the author's will, and since these styles cannot be consciously manipulated by the author, they are considered to provide the most reliable data for a Stylometry study[3].Recently, it has gained greater importance due to its applications in forensics analysis, humanities and electronic commerce. Stylometric analysis is important to marketers, analysts and social sciencetist.

Research into authorship attribution is going on from 19th Century but it determines authorship of document with more than 1000 words. By 1990's this value is decreased to 200 words. In 21st Century it is possible to find determine authorship of a document with 250 words.

Due to increases in usage of emails for communication as well as for attempting crimes, Author Identification can be applied for performing forensic analysis of online messages in cybercrime investigation.

First section will explain the concept of n-gram. Second section proposes the use of n-gram in combination with Stylometry followed by experimental results on Reuter_50_50 Data set .

II. Methodology

A. N-gram (Character /word n-gram):

In many established approaches to Stylometry, the (relative) frequencies of the most frequent words (MFW) in a corpus are used as the basis for multidimensional analyses. Along with that other features are also worth considering, especially word and/or character n-grams. The idea behind use of n-grams is to combine a string of individual items into partially overlapping, consecutive sequences of n of these individual items.

Given a sample sentence \This is a simple example", the character 2-grams (\bigrams") are as follows: \th", \hi", \is", \s ", \ i", \is", \s ", \ a", \a ",\ s", \si", \im", \mp", etc. The same sentence split into bigrams of words reads \this is", \is a", \a simple", \simple example".

B. Stylometry Approach

The various steps for calculating Authorship is as shown in figure1, Step1 is calculating frequencies of words and identifying most frequent words from entire corpus. Then step2 calculates Normalized frequency by calculating total percentage of the most frequent word in that document with respect to entire corpus. In step3 Z-score approach is used.Step4 calculates distance table by finding distance between two matrices. Thus by representing text into numeric representation that is feature extraction, clustering and classification techniques of machine learning can be applied.

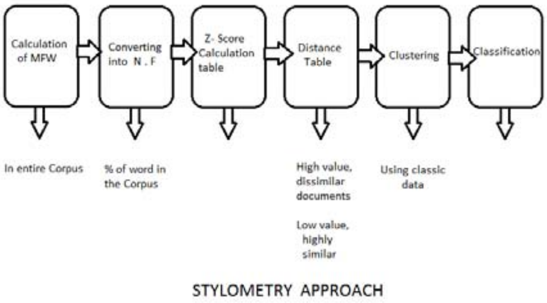

Fig.1 Steps for Authorship Identification

III. Experimental Methodology

A. Feature Extraction

It processes electronic texts to create a list of all the words used in all texts of corpus, with their frequencies in the individual texts; normalizes the frequencies with z-scores (if applicable).additional procedures that (usually) improve attribution are

1. Automatic deletion of personal pronouns and
2. Culling (automatic removal of words too characteristic for individual texts)

To explain the process of feature extraction, we are considering 4 authors and 6 frequent words as shown below.

Table1 is showing sample of Normalized frequency table with 6 frequent words for 4 authors named as A1, A2, A3, and A4.Mean and standard deviation is calculated for every word.

	W1	W2	W3	W4	W5	W6
A1	2.675	2.551	1.673	1.993	2.107	1.942
A2	3.284	2.996	2.7	2.1	1.706	1.49
A3	2.852	2.721	2.5	2.58	1.59	1.96
A4	2.608	3.048	1.0	1.607	1.94	1.30
Mean	2.85	2.829	1.99	2.077	1.83	1.679
Standard Deviation	0.30	0.234	0.75	0.403	0.23	0.328

Table.1 Normalized Frequency Table

The values in table1 are converted into “Z-Score”, which reflect the extent to which the normalized word frequencies within a particular text are above or below average for the set of text as a whole. If they are negative it indicates that text is below average. If they are negative it indicates that text is below average. The formula used for calculating Z-Score is as follows.

z-score= (NF-mean)/SD (1)

Table 2 shows Z-score for corresponding normalized frequency table.

	W1	W2	W3	W4	W5	W6
A1	-0.592	-1.187	-0.412	-0.207	1.167	0.80
A2	1.412	0.713	0.925	0.115	-0.56	-0.55
A3	-0.009	-0.459	0.703	1.259	-1.055	0.887
A4	-0.810	0.933	-1.216	-1.166	0.452	-1.138

Table.2 z-score Table

The values in Table3 shows Distance table which is calculated from z-score of table2 by applying delta distance formula.

	A1	A2	A3	A4
A1	0	1.441	1.033	1.125
A2	1.441	0	0.981	1.243
A3	1.033	0.981	0	1.676
A4	1.125	1.243	1.676	0

Table3. Distance Table

From distance table we can conclude that A1 and A4 are consider as similar text and can be represented in one cluster as distance is less and A2 and A3 are closely associated as distance is less.

IV. Experimental Evaluation

Reuter_50_50 Data set is used for experiments. It consists of total 50 authors and 50 documents per author. Therefore training corpus consists of 2,500 texts and test corpus also consists of 2500 text which is non-overlapping with training texts.

Dataset	Measures	Delta	KNN	SVM
		Stylometry Features	Stylometry Features	Stylometry Features
Reuter_50_50 Data set For bi-grams	Avg. Accuracy	65%	60%	77%
Reuter_50_50 Data set For unigram	Avg. Accuracy	67%	69.23%	85.01%

Table.4 Experimental Results

V. Conclusion

By introducing Stylometry approach and n-gram features in the task of authorship identification we have achieved 85% of accuracy for SVM classifier which is more in compared to Delta and KNN classifier. Dataset used was Reuter_50_50 Data set. Also, we have compared Unigram and bi-gram approach for all the three classifiers. In our future work, the accuracy of the classification can be improved by finding and incorporating more features.

References

[1] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott,”Stylometry for E-mail Author Identification and Authentication”,Seidenberg School of CSIS, Pace University, New York

[2] Ahmed Abbasi and Hsinchun Chen,”Writeprints: A stylometric approach to identity-level identi_cation and similarity detection in cyberspace”, ACM Trans.Inf. Syst., 26(2):1{29, 2008.

[3]. Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelg an ger _nder: Taking stylometry to the underground. 2014.

[4]. Shlomo Argamon, Moshe Koppel, and Galit Avneri. Routing documents accordingto style. In First International workshop on innovative information systems, pages 85{92. Citeseer, 1998.

[5]. Shlomo Argamon, Marin _Sari_c, and Sterling S Stein. Style mining of electronicmessages for multiple authorship discrimination: _rst results. In Proceedings of theninth ACM SIGKDD international conference on Knowledge discovery and datamining, pages 475{480. ACM, 2003.

[6]. Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In 6th JADT, pages 29{37. Citeseer, 2002.

[7] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3):121{132, 1996.

[8]. Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In Big Data Analytics, pages 37{47. Springer,2013.