IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 04 (April. 2014), ||V2|| PP 27-30

Use of Stylometry and Outlier Detection Algorithm in Online Writing Sample to Detect Outliers

Sonia Sharma, Pushpendra Kumar Petrya
[1](M.Tech Scholar, Department of Computer Science and Engg, Lovely Professional University (LPU))
[2](Assistant Professor, Department of Computer Science and Engg, Lovely Professional University)

Abstract: - Outlier Detection, now a days, is one of the emerging technology used in data mining. The data objects which deviate from the other data objects in the data set are considered to be as outliers. Outliers are classified as global outliers, collective outliers and contextual outliers. Outlier detection contains a broad spectrum of techniques to detect outliers.[1] Here, we are going to propose an algorithm which detects outliers (unmatched sample) in online writing sample. The online writing sample is analysed firstly by using the well known concept 'Stylometry'. Stylometry is the study which helps to distinguish between the writing style of two persons. It is assumed that the writing style of two persons always differs and every person contains a feature in the sample which defines the uniqueness of the individual.[2] This model can be used in detecting plagiarism, e-mail verification and author identification. The idea behind this proposed model is to save the data from any criminal activities so that the correct author should be identified.

Keywords: - *Feature Selection, Outlier detection, Stylometry, Threshold metrics, Unexpected parameter*

I. INTRODUCTION

Data mining is one of the fast growing field now a days. It implies extraction of hidden information from the large datasets. Various techniques are used in the data mining. Outlier detection is one of the important technique used in the data mining.

1.1. Outlier Detection

Outlier detection, as we said, is one of the important branch in the data mining. It is the process by which we can find the data objects which have different behavior than the other data objects. The various outlier detection methods are Supervised methods, Unsupervised methods, Semi-Supervised methods, Statistical methods, Proximity-based methods, Clustering-based methods[7]. Outlier detection has many applications in the field of credit card fraud detection, security, image processing, medical field and intrusion detection.[1]

1.2 Stylometry

It is the study in which by the writing style a person can judge another person. It is the study of the writing behavior of the person. It is mainly used to determine the writing style of a person. The analysis is done based on the various features. The features include usage of vocabulary, usage of function words, structure and length of sentences and formatting of text on the page. It is also assumed that the writing style and the features of person remain constant throughout all his writings. In today's education system, plagiarism is an serious issue. Stylometry can also be used to detect the plagiarism.
So, In this paper, we purposed a model which will detect the outliers (unmatched sample) in any online samples by using stylometry to analyze the writing sample and then an algorithm.

II. RELATED WORK

Ramyaa et.al defines a model, "Using machine learning techniques for stylometry" to justify the use of various techniques in stylometry. The results were obtained using Decision trees and neural networks on the test set.[3] D.Pavelec et.al proposed another model, "Compression and Stylometry for Author Identification" which uses two different paradigms for author identification. The one paradigm is compression based algorithm and the second is classical pattern recognition framework.[4] Robert Goodman et.al proposed "The use of Stylometry for Email Author identification: A Feasibility Study". They applied a stylometric analysis of biometric keystrokes for authorship attribution.[5] Oral Alan et.al proposed "An Outlier Detection Algorithm Based on Object-Oriented Metrices Thresholds" in dataset based on some threshold metrices.[6]

III. PROPOSED WORK

This paper includes the work in Stylometry and Outlier detection. The analysis of handwriting patterns has to be done to detect any unmatched pattern in the sample of an individual. The analysis of the documents

has to be done using Stylometry in which writing pattern of an individual has been analysed based on the different writing features. Then the parameter(s) should be selected to claim the uniqueness of the individual. If there is any unexpected behavior analysed in the writing sample that should be considered as an outlier, and is removed using an outlier detection algorithm.

This scheme introduces three phases, the first phase includes the analysis of the writing sample of the individual using the concept of Stylometry, the second phase includes the selection of parameter which is more frequently used by the individual, The third phase includes the detection of the unexpected pattern in the writing sample which is considered as an outlier and then detection of that outlier using an outlier detection algorithm, which includes the comparison of predefined threshold values and parameter.
The process of work flow is illustrated in the Fig. 1.

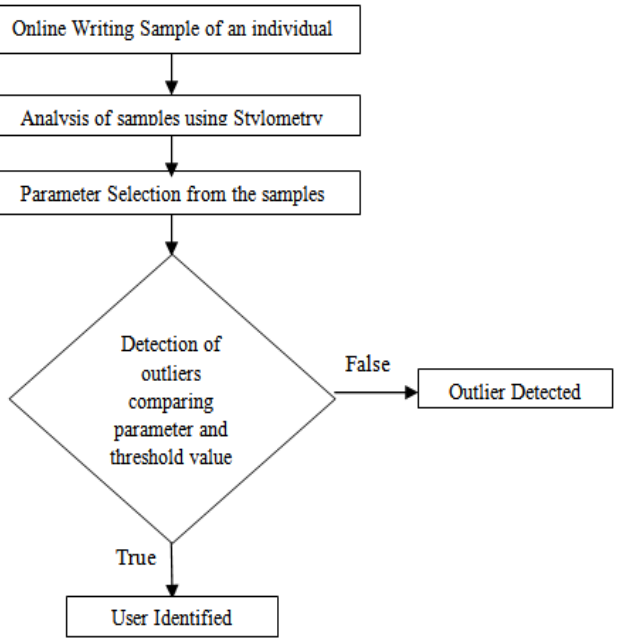

Fig 1. Proposed Work

3.1 Phase Description
There are separate phases in this work, which are essential for the effective working of the process, The phases has been explained below

3.1.1. Analysis Phase

In this the analysis of the features to be done based on some linguistic features. Till now, In the stylometry the study has been done on sixty two features including the tagged text, Parsed text, Interpreted text[3] e.g. , Number of words, Number of Sentences, Number of commas, Number of dashes, Number of Semi-colons, number of Question-marks, Number of underscores. These are examples of some features used till now . Three more features have been included in this paper. The features are

- Number of *well* per hundred token
- Number of *particularly* per hundred token
- Number of *must* per hundred token

3.1.2. Parameter Selection Phase

The next phase includes selection of an appropriate parameter which identifies the uniqueness of a writing sample of an individual. Thus the pre-defined threshold frequency range should be provided to each parameter used for analysis. The parameter should be selected based on the frequencies given.

3.1.3. Detection of Outliers.

The next phase includes the detection of unmatched pattern or unexpected parameter behaviour in the writing sample, thus that is considered as an outlier and is identified using an outlier detection algorithm which is based on the threshold defined for the metrics for the dataset. If the value of parameter does not match the threshold value of parameter, then it is considered to be as outlier. The certain metrices need to be defined and also their threshold values.

Metric	Abbreviation	Metric	Abbreviation
Number of commas per hundred tokens	CPHT	Number of semicolon per hundred tokens	SPHT
Number of multiple commas per hundred tokens	MCPHT	Number of *the* per hundred token	TPHT
Number of dashes per hundred tokens	DPHT	Number of *well* per hundred token	WPHT
Number of Exclamation sign per hundred tokens	EPHT	Number of *particularly* per hundred token	PPHT
Number of question mark per hundred tokens	QPHT	Number of *must* per hundred token	MPHT

Table 1. Metrices and Abbreviations

The Outlier detection algorithm is given as

Outlier detection algorithm (*Detection of outliers in the sample dataset.*)
Inputs Writing Sample/ New Writing Sample.
Output User Identified/ Detection of Outlier.
Initialization of THR_CPHT, THR_MCPHT, THR_DPHT, THR_EPHT, THR_QPHT, THR_SPHT, THR_TPHT, THR_WPHT, THR_PPHT,
for *parameter* in *DATASET*

if(*parameter*.CPHT!=THR_CPHT),
sample data is an outlier,
end if
if(*parameter*.MCPHT!=THR_MCPHT),
sample data is an outlier,
end if
if(*parameter*.DPHT!=THR_DPHT),
sample data is an outlier,
end if
if(*parameter*.EPHT!=THR_EPHT),
sample data is an outlier,
end if
if(*parameter*.QPHT!=THR_QPHT),
sample data is an outlier,
end if
if(*parameter*.SPHT!=THR_SPHT),
sample data is an outlier,
end if
if(*parameter*.TPHT!=THR_TPHT),
sample data is an outlier,
end if
if(*parameter*.WPHT!=THR_WPHT),
sample data is an outlier,
end if
if(*parameter*.PPHT!=THR_PPHT),
sample data is an outlier,
end if
if(*parameter*.MPHT!=THR_MPHT),
sample data is an outlier,
end if
else
Sample data is correct

end for
END Algorithm

Table 2. Outlier Detection Algorithm

IV. CONCLUSION

In this paper, a framework for the detection of unexpected parameters in online samples has been introduced. It uses the technique of stylometry for the analysis of the sample and then a feature selection phase has also been included, Which helps to identify the parameter which is most frequently used by the person and thus claim the individuality of the person. The next phase includes the detection of the unmatched pattern using an outlier detection algorithm , which includes the comparison of pre-defined threshold values with the parameters. Thus if the value of parameter does not match the threshold value, then it is an outlier. In future, more features can be added and thus analysis can be done based on that features.

REFERENCES

[1] Jiwani Hen, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques",Outlier Detection, pp No. 546-576

[2] *Stylometry*. (n.d.). Retrieved December 2013, from wikipedia: http://en.wikipedia.org/wiki/Stylometry

[3] *Supervised Outlier Detection*. (n.d.). Retrieved December 2013, from link.springer.com: http://link.springer.com/chapter/10.1007%2F978-1-4614-6396-2_6#page-1

[4] Ramyaa, He Congzhou, Rasheed Khaleed, "Using machine learning techniques for stylometry", Artificial intelligence Center, The university of Georgia, Athenes.

[5] Pavelec D, Oliviera L.S, Justino E, Neto F.D Nobre, Batista L.V, "Compression and Stylometry for Author identification", Curitiba, Brazil.

[6] Goodman Robert, Hahn Matthew, Marella Madhuri, Ojar Christina, Westcott Sandy(2007), "The use of Stylometry for Email Author identification: A feasibility study", Seidenberg School of CSIS, Pace University,1 Martine Ave, White Plains, NY, 10606, USA.

[7] Alan Oral, Catal Cagatay(2009), "An Outlier detection algorithm based on the object oriented metrics threshold", Tubitak-Marmara research center, Information Technologies Institute, Turkey.