Authorship Verification

Nilesh Solanki
solank01@ads.uni-passau.de
University Of Passau

Smruti Ranjan Mohapatra
mohapa01@ads.uni-passau.de
University Of Passau

Rohan Sapate
sapate01@ads.uni-passau.de
University Of Passau

1 INTRODUCTION

Authorship analysis is the process of examining the characteristics of a piece of work in order to draw conclusions on its authorship. Authorship analysis is the growing problem in this digitalized world and there has been an extensive study in its field. There are three subfields of authorship analysis: Authorship Profiling, Authorship Identification and Authorship Verification[1]. In the cybercrime world, the culprits are unidentified and thus the task of the forensic experts is to find their age, gender, nationality and other personal details (-in some cases demographics) from the texts they have written which is called authorship profiling. There are a lot of unauthored documents which a lot of people are falsely claiming to be their own. We can solve this problem and find the most likely author of the unknown document when provided with some existing texts from respective authors, which is called authorship identification, which is the second scenario. And the third scenario, and on which we will be working is called authorship verification. The task of authorship verification is to find out the legitimacy and ownership of the given text as to whether it is written by the same author who has written other texts in our database. The answer should be yes or no. Our emphasis is to discover more details, problems and various approaches required to address authorship verification. In this project we will work on understanding different methodologies used for this task like Common N-Gram profiles of text documents and apply to detect whether the document is written by the author. The various authorship attribution methods can be considered in two paradigms:

Profile based Paradigm: Unlike instance based paradigm, all possible text samples of the possible author are treated together. This is done by merging them into one large document and they are rendered into becoming a single profile of the author.

Instance based paradigm: where the samples of the author are considered and treated individually. Considering if there is only one document accessible for a possible author then splitting of the document into multiple samples is carried out.

Our primary focus will be on using instance based paradigm which is beneficial when long documents that can be splitted into samples are to be considered. To back up our literature, in the PAN 2013 evaluation campaign, the most widely preferred paradigm was instance based. A total number of 17 participants out of 18 favoured instance based paradigm and found out that the instance based paradigm was the most fitting one [2]. In this project we will try to apply the knowledge of text processing, tokenization, segmentation, frequency distribution, conditional frequency distribution, n-gram.

2 TARGET PLANS

In this project, we will be using the algorithm based on dissimilarity measure as mentioned in (Magdalena Jankowska et al, 2013). In this algorithm, it is initially assumed that we are given a set of known documents and an unknown document. Our main task is to detect whether the author wrote it or not. Hence the result should be Yes(1) or No(0). For each known document, we would find the maximum dissimilarity using dissimilarity measure between the considered document and rest of the known documents. Likewise we also evaluate dissimilarity between each known document and unknown document.Then we would find the ratio of each dissimilarity of the unknown and known documents to the maximum dissimilarity found for each document. And we would take average of them. This average would be subjected to a threshold value which later on determine the authorship of the unknown document.

REFERENCES

[1] Sara Elmanarelbouanani and Ismail Kassou. 2013. Authorship Analysis Studies: A Survey. *International Journal of Computer Applications* 86 (12 2013). https://doi.org/10.5120/15038-3384

[2] Patrick Juola and Efstathios Stamatatos. 2013. Overview of the author identification task at PAN 2013. *CEUR Workshop Proceedings* 1179 (01 2013).

Text Mining Lab, 2020, Uni Passau
.