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 two texts in order to draw conclusions if they are written by the same author or not. It is the growing problem in this digitalized world and there has been an extensive study in this area. There are three subfields of authorship analysis: Authorship Profiling, Authorship Identification and Authorship Verification[4] 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 if the given documents are written by the same author. Our emphasis is to discover more details, problems and various approaches required to address authorship verification. Every author has their own writing style, so we can distinguish a particular author from other authors based on the writing style [6]. 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.

Authorship Verification is one of the tasks from PAN 2020. In this task, in the given data, there are pairs of texts given and the objective of the task is to determine if both the documents in each pair are written by the same author or not. In this project we will try to apply the knowledge of text processing, tokenization, segmentation, frequency distribution, conditional frequency distribution, n-gram and so on.

2 PLANS

In PAN 2020 challenge, we have baseline code which uses cosine similarity method with TFIDF-normalized, bag-of-character-tetragrams as a feature. Also, we have the evaluation code which comprises of four evaluators namely : F1, AUC, c@1 and F_0.5u. In our experiment, we will compare our methodology as described in the following section with the baseline method of PAN 2020 challenge. Our target is to improve the evaluation score of the baseline method

Text Mining Lab, 2020, Uni Passau
.

3 FEATURES

For author verification task, a number of feature sets have already been used. In Koppel and Winter, character-tetragrams representations were used as a feature in similarity based methods[9]. In Maitra et. al, 8 different feature sets were used like total punctuation count, specific punctuation ratio, long-sentence or short-sentence ratio, vocabulary strength, N-gram difference, POS frequency, POS sequence frequency and starting POS Frequency to evaluate in Random Forest based method[10]. Koppel and Schler used 250 most frequent word uni gram in unmasking method for authorship verification[8]. In authorship attribution experiments, in Koppel et. al POS tagging was used [8] and in Baayen et al punctuation marks and function words were used [2]. For PAN2019 attribution tasks, the most common features used by majority participants in the task which provided sound results were character n grams, word n grams, distortion, POS, punctuation, token ,etc.[7]. In Castro et al, a verification task with average similarity method there were a number of features considered which can be categorised as character n grams , word n grams and Lemma and POS[5]. Our method is inspired from Jankowska et al, where character n-grams and word n-grams were used as features.[5] In our experiment, we would like to add some more features like POS frequency, punctuation counts, vocabulary strength in addition to character n gram and word n gram.

4 METHODOLOGY

In this project, we will be using the algorithm based on dissimilarity measure as mentioned in [5]. 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 average of all ratio calculated by dividing each dissimilarity of the unknown and known documents to the maximum dissimilarity found for each document and we term it as average dissimilarity ratio that is subjected to thresholding value which later on determine the authorship of the unknown document.

There are two datasets available from PAN 2020: a small dataset and a large dataset. Our primary focus is on the smaller dataset. This dataset is in .jsonl format and comprises of two files, in which the first file contains pair of text where each pair has a unique id and the second file is a ground truth for all pairs which indicates if the text in the pair from the first file are from the same author or not [1]. We will extract the mentioned features like character n-gram' and word n-grams from the NLTK package. Similarly for the feature POS frequency we will use taggers from NLTK such as Default Tagger, Unigram Tagger, Bigram Tagger. For the punctuation feature,we would be employing regular expressions to find out all punctuations. Another feature namely vocabulary strength can be calculated from the unique words present in the text and total number of words present in the text[10]. For each feature, we will be employing an algorithm with dissimilarity measures as mentioned above to detect whether the given text is written by the given author or not and we will be following the evaluation procedure as mentioned below.

5 EVALUATION

The evaluation code is provided by PAN which will evaluate the project based on four parameters:

- **AUC :** Area under the curve, which is the measure of separability. So higher AUC means the model is better at predicting the 0s as 0s and 1s as 1s [11] papers.
- **F1-score :** F1-score of a model is the harmonic mean of precision and recall. As both precision and recall are important to determine the accuracy of any model.
- **C@1 :** This measure prefers not responding than responding incorrectly [12]. It means, out of two models, this measure will give more preference to the model that has the same correct answer but less incorrect answer by choosing not to answer some questions.
- **F_0.5u :** It is modified version of F0.5 where the non answers are treated as false negatives which puts more emphasis on deciding same-author cases correctly[1] [3].

REFERENCES

[1] [n.d.]. *Online reference about the section.* https://pan.webis.de/clef20/pan20-web/author-identification.html

[2] Harald Baayen, Hans Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. *IEEE Intelligent Systems Their Applications - IEEE INTELL SYST APPL.*

[3] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing Unmasking for Short Texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 654–659. https://doi.org/10.18653/v1/N19-1068

[4] 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

[5] Magdalena Jankowska, Evangelos Milios, and Vlado Keselj. 2014. Author Verification Using Common N-Gram Profiles of Text Documents. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (Aug. 2014), 387–397. https://www.aclweb.org/anthology/C14-1038.pdf

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

[7] Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-Domain Authorship Attribution Task at PAN 2019. (Sept. 2019). http://ceur-ws.org/Vol-2380/paper_264.pdf

[8] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational Methods in Authorship Attribution. *JASIST* 60 (01 2009), 9–26. https://doi.org/10.1002/asi.20961

[9] Moshe Koppel and Yaron Winter. 2014. Determining If Two Documents Are Written by the Same Author. *Journal of the Association for Information Science and Technology* 65 (01 2014). https://doi.org/10.1002/asi.22954

[10] Promita Maitra, Souvick Ghosh, and Dipankar Das. 2016. Authorship Verification - An Approach based on Random Forest. (07 2016).

[11] Sarang Narkhede. [n.d.]. Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[12] Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-response., Vol. 1. 1415–1424.