A Multi-Feature Custom Classification Approach to Authorship Verification

Notebook for PAN at CLEF 2023

Riya Sanjesh [1] and Alamelu Mangai [2]

[1] *Presidency University, Ittagallpura, Bengaluru, India*
[2] *Presidency University, Ittagallpura, Bengaluru, India*

Abstract

PAN 2023 conducted Cross Discourse Authorship Verification task to evaluate the systems that identify if a pair of text was authored by the same person. This paper talks about our approach towards this task. We used multiple stylometric features such as character n-grams, POS tags, character and word count etc. to represent the texts. The predictions were done using a custom classifier which uses cosine similarities between the text of an author and the input pair of texts. This submission of ours achieved the AUC of **0.985** on the training data set.

Keywords

Authorship Verification, PAN 2023, Multi-Feature extraction, Stylometric Features, Cosine Similarity

1. Introduction

Authorship Verification task [1, 2] is being conducted by PAN at CLEF since 2013 [3, 4, 5, 6, 7, 8]. This task aims to determine if a pair of text has been written by the same author. In this paper we introduce our work on Cross-Discourse Type Authorship Verification task in PAN at CLEF 2023 [9]. In this edition of the Authorship Verification task the focus was on cross discourse type authorship verification where both written and spoken languages were represented in the set of discourse types.

In our proposed system we introduced a multi-feature extraction and transformation to represent the texts and a custom classifier based on cosine similarities to make the predictions.

2. Proposed Approach

In this section we talk about the data set used and describe our approach in detail.

2.1. Data set

PAN 2023 Cross Discourse Authorship Verification task provided training (calibration) and test data set to train and evaluate the submitted tasks. The data consists of pairs of texts (both in English) from two different discourse types and given a unique identifier. The structure of both training and test data set is the same. The training data contains 8836 pairs along with the metadata on the discourse type for each pair of text. We divided this dataset into two parts – one for training purpose and the other for validating the proposed model. The ratio between the training and validation set was 80 – 20. Table 1 shows the breakdown of the data sets.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece
EMAIL: riya.20223CSE0038@presidencyuniversity.in (Riya Sanjesh); alamelu.jothidurai@presidencyuniversity.in (Alamelu Mangai);

Table 1
Training and Validation data set sizes

	Training Set	Validation Set
Number of pairs	7068	1768

2.2. Proposed System

To train the system we used 3 stages through which we pass the texts. The first stage preprocesses the texts using NLTK packages to extract the tokens and POS tags. This stage is followed by Feature extraction which extracts different types of features described in the next sub section. Once the texts are transformed and fitted into the model we move to the next stage where we measure the cosine similarities of the pairs of texts and with the text of individual authors. We believe that if a pair of text was written by a single author then both these text should be similar to the other texts written by that author and not just similar to each other. Due to this reason we combined all the texts written by an author into a single text for this comparison. This data is further taken through a custom classifier to come up with best configuration to predict the outcomes. All these stages are described in the below sections.

2.2.1. Preprocessing and Feature Extraction

All the text from a single author is clubbed together into a single document. These documents are further passed to the following stages. The first stage of our proposed systems consists of a preprocessor and a multi-feature transformation to extract different types of features from the input text. The preprocessor uses NLTK Treebank Word Tokenizer to tokenize the text and NLTK's POS tagger to POS tag.

We used multiple features which are extracted from the input data set. Some of these features are adopted from [10] and some of them also described in Writeprints feature set [11]. Following are the features that we used:

- TFIDF for character n-grams where n is between 1-6
- TFIDF for n-grams of POS tags where n is between 1-3
- TFIDF for n-grams of POS tag chunks where n is between 1-3
- TFIDF for punctuation marks used in the text
- Frequency of stop words
- Count of characters in the text
- Count of words in the text
- Average character counts per word
- Ratio of hapax-legomenon and dis-legomenon

2.2.2. Training the classifiers

We worked on two different classifiers with slight differences. The motivation behind the second classifier was to understand the impact of leaving some the problems unanswered in the borderline cases. Below we describe each one of them.

2.2.2.1. Classifier 1

Our custom classifier calculates the similarities of each pair of texts among themselves as well as with the individual authors' texts. This gives us three values for each author and pair of texts. For example, consider a case with three authors, A, B, C and a pair of text, say x1 and x2. Our classifier

calculates the similarities between x1 and x2, A and x1, A and x2, B and x1, B and x2, C and x1 and C and x2 as shown in the Table 2. This is repeated for each pair of text. For calculating the similarities we use cosine similarity.

Table 2
Cosine Similarities

Author	Distance of x1 to Author	Distance of x2 to Author	Distance between x1 and x2
A	A->x1	A->x2	x1->x2
B	B->x1	B->x2	x1->x2
C	C->x1	C->x2	x1->x2

For each pair of text we get the number of rows in the above table, equal to the number of authors in the input data set. We believe that all these three values are important and due to this our classifier tries to maximize these values. To achieve this, we identify a row which has the max value for the summation of all three distances, for each input text pair. We store all three values of such a row, let's say, s1, s2 and s3. Now we calculate a cutoff value for each column based on these values and a threshold value as mentioned below:

$$\text{cutoff1} = \text{thr} * \text{s1} \quad (1)$$
$$\text{cutoff2} = \text{thr} * \text{s2} \quad (2)$$
$$\text{cutoff3} = \text{thr} * \text{s3} \quad (3)$$

where, *thr* is the threshold value. The purpose of the cutoff is to find out how far ahead are the best values for the cosine similarities from the rest of the values (cosine similarities of the pair of the text and the text from other authors).

In the next step, we try to find all rows for a pair with values above the cutoff values. If we get exactly one such row, we predict the pair written by the same author. In all other cases we predict the authors of the pair to be different.

This process is repeated for threshold values ranging from 0.2 to 1 to identify the value that gives us the best values for our predictions.

1.1.1.1. Classifier 2

Classifier 2 is exactly the same as the classifier 1 with just one difference in the prediction. Classifier 2 leaves the problem unanswered if we get exactly two rows above the cutoff values. Classifier 1 in this case, predicts the authors of the pair to be different.

1.1.2. Evaluation and Results

PAN 2023 evaluation set is used to evaluate the model. This evaluation uses 5 metrics against which the submissions are compared and ranked. While training our proposed model we used the 'overall' score which is the mean of all the 5 metrics used by the evaluator. Table 3 shows the results of our training runs for different threshold values ranging from 0.2 to 0.95 at step size of 0.05.

Table 3
Various threshold values and corresponding overall score

Threshold	Overall Score
0.2	0.677
0.25	0.677
0.3	0.677

0.35	0.678
0.4	0.686
0.45	0.712
0.5	0.742
0.55	0.759
0.6	0.772
0.65	0.779
0.7	**0.783**
0.75	0.776
0.8	0.77
0.85	0.767
0.9	0.767
0.95	0.766

We further tried to fine tune the threshold value by repeating the training process for threshold values ranging from 0.65 to 0.8 and step size of 0.01 and best value for threshold was chosen to be **0.69** which resulted in the best overall score of **0.785**. We took this threshold and used it for the run of our system on the validation data set. Table 4 shows the overall score on both training and validation data sets.

Table 4
Results of our system on the training and validation data sets at threshold 0.69

Classifier	Training set	Validation set
Classifier 1	0.785	0.763
Classifier 2	0.786	0.763

The proposed system with Classifier 1 (calm-lyrics) and Classifier 2 (null-midpoint) were then submitted on TIRA Platform [12] for PAN 2023 Cross Discourse Authorship Verification task[2] and Table 5 shows the scores for the 5 metrics when the system was run on the training data provided by the platform.

Table 5
Results from PAN 2023 Cross Discourse Authorship Verification[1]

Classifier	AUC	C@1	F_05_U	F1	Brier	Overall
classifier 1	0.985	0.781	0.745	0.815	0.847	0.835
classifier 2	0.984	0.783	0.742	0.82	0.845	0.835

2. Conclusion

The proposed system performed fairly in PAN 2023 Cross Discourse Authorship Verification task on the training data set. Although Classifier 2 got a slightly better C@1 and F1 score as it leaves some of the problems unanswered but the overall score was same as that of the classifier 1. As a future work we would like to look at the multi-feature key extraction more closely and analyse the impact of different combinations of features. We would also like to look at how we can improve our predictions when the authors of the test data set are unknown.

[2] https://pan.webis.de/clef23/pan23-web/author-identification.html

3. References

[1] O. Halvani, L. Graner, R. Regev, Taveer: an interpretable topic-agnostic authorship verification method, in: M. Volkamer, C. Wressnegger (Eds.), ARES 2020: The 15th International Conference on Availability, Reliability and Security, ACM, 2020, pp. 41:1–41:10.

[2] N. Potha, E. Stamatatos, Improving author verification based on topic modeling, Journal of the Association for Information Science and Technology 70 (2019) 1074–1088.

[3] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E.Stamatatos, B.Stein, M. Wiegmann,M.Wolska, E.Zangerle. Overview of PAN 2022: AuthorshipVerification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. B. Cedeno, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Proceedings of theThirteenth International Conference of the CLEF Association (CLEF 2022), Springer, 2022.

[4] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, M. Potthast, B. Stein. Overview of the Authorship Verification Task at PAN2022. Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2022).

[5] Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, CLEF 2020 Labs and Workshops, Notebook Papers, September 2020. CEUR-WS.org.

[6] Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2018). Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al. (pp. 1-25).

[7] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the Author Identification Task at PAN 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, Working Notes Papers of the CLEF 2014 Evaluation Labs, volume 1180 of Lecture Notes in Computer Science, September 2014.

[8] Patrick Juola and Efstathios Stamatatos. Overview of the Author Identification Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, September 2013. CEUR-WS.org

[9] Martin Potthast, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Cross-Discourse Type Authorship Verification at PAN 2023, in: CLEF 2023 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2023

[10] Janith Weerasinghe and Rachel Greenstadt. "Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification" Notebook for PAN at CLEF 2020

[11] Abbasi, A.. Chen, H.c.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems 26, 1{29 (01 2008)}.

[12] M. Froebe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, Springer, 2023, pp. 231–241