UniNE at PAN-CLEF 2020: Author Verification
Notebook for PAN at CLEF 2020

Catherine Ikae[1]

Computer Science Department, University of Neuchatel, Switzerland
Catherine.Ikae@unine.ch

Abstract. In our participation in the authorship verification task (small corpus), our main objective is to be able to discriminate between pairs of texts that were written by the same author (denoted "same-author") and pairs of snippets written by different ones ("different-authors"). The paper describes a simple model that performs this task based on a Labbé similarity. As features, we employed the most frequent tokens (words, and punctuation symbols) from each author after including the most frequent ones of a given language. Such a representation strategy is based on word used frequently by a given author but not belonging to the most frequent in the English language. Evaluation based of authorship verification task with a rather small set of features shows an overall performance with the small dataset of F1= 0.705 and AUC = 0.840.

1 Introduction

Authorship verification at CLEF PAN 2020 is the task of determining whether two texts (or excerpts) have been written by the same author [1]. In this kind of task, one can also provide a sample of texts written by the proposed author from which the system could generate a better author's profile. This additional sample was not provided in the current experiment.

For the CLEF PAN 2020 (small corpus), the pairs of texts have been extracted from the website `www.fanfiction.net` storing texts about numerous well-known novels, movies or TV series (e.g., *Harry Potter, Twilight, Bible*)[2]. These writings called fanfics have been scripted not by the true author(s) but by fans who want to continue or propose a new episode for their preferred saga. Of course, such a fan could have written for different series or propose several variants for continuing a specific one. One can however assume that a writer is more interested to script in a given topic or domain (called fandom, a subculture of fans sharing a common interest on a given subject).

Thus for the proposed task, the question is to identify whether or not a pair of text excerpts have been authored by the same person. We view this author verification as a

similarity detection problem, or to detect when a similarity between two texts is too high to reflect two distinct authors.

Just like in authorship attribution, the author of a given text had to be revealed by identifying some of his/her stylistic idiosyncrasies and to measure the similarity between two author's profile. [3] Suggest using quantification of writing style in texts to represent the identity of their authors. [4] Makes use of emojis in the feature selection for verification of twitter users. [5] Applies large numbers of linguistic information such as vocabulary, lexical patterns, syntax, semantics, information content, and item distribution through a text for author recognition and verification.

As possible application of author verification, one can mention analysis of anonymous emails for forensic investigations [6], verification of historical literature [7], continuous authentication used in cybersecurity [8], detection of changes in writing styles with Alzheimer patients [9].

The rest of this paper is organized as follows. Section 2 describes the text datasets while Section 3 describes the features used for the classification. Section 4 explains the similarity measure and Section 5 depicts some of our evaluation results. A conclusion draws the main findings of our experiments.

2 Corpus

The corpus contains a set of pairs composed of two short texts (or snippets) describing proposed variants or continuation of a series achieving a popular success. In this context, given two snippets, the task is to determine whether the text pair has been written by a single author or by two distinct writers. This question is a stylistic similarity detection problem assuming that two snippets could cover distinct topics with very distinctive characters and temporal differences but still being written by the same person.

These pairs of texts have been extracted from different domains (fandoms) and Table 1 reports some examples of such fandoms with the number of texts extracted from them.

('G-Gundam', 56), ('Vampire Knight', 175), ('Free! - Iwatobi Swim Club', 121), ('DC Superheroes', 98), ('Friends', 111), ('CSI: Miami', 133), ('Grimm', 108), ('Danny Phantom', 200), ('Primeval', 117), ('Kingdom Hearts', 219), ('Jurassic Park', 107), ('Tarzan', 40), ('Dungeons and Dragons', 73), ('Final Fantasy X', 152), ('Fast and the Furious', 112), ('OZ', 33), ('Sons of Anarchy', 115), ('Avatar: Last Airbender', 223), ('Attack on Titan', 194), ('Madam Secretary', 47)

Table 1: Examples sample fandom and number of times they appear in the pairs.

This corpus contains 52,590 text pairs (denoted problems) from which 27,823 pairs corresponding to the same author and 24,767 are pairs written by two distinct persons. Each text excerpt contains, in mean, 2,200 word-tokens. An example of a pair is provided in Table 2 with their respective length and their vocabulary size.

Guardians of Ga'Hoole	Tokens = 2,235	| Voc |= 1,353
I shift a bit, warily letting my eyes dart from one owl to the other -- but my eyes are trained on the Barn Owl the most. Like Hoole...so like Hoole... He turns a bit, and our eyes meet directly. I can"t describe it...in this next moment, I don"t look away, how awkward it seems. I stare into his eyes. They"re like Hoole"s... They are Barn Owl eyes, but Hoole"s eyes. They"re his eyes...Hoole"s eyes... They hold that light of valor, …		
Hetalia - Axis Powers	Tokens = 2,032	| Voc |= 1,422
"All will become one with Russia," he said, almost simply, his cheer eerie. Fists were already clenched; now they groped about, for a pan, a rifle, a sword-there was nothing. In some way, this brought her but a sigh of relief-Gilbert and Roderich, she was reminded, were not here to suffer as well. If Ivan put his giant hands on Roderich... Click, went an object, and Elizaveta was snapped into the world when her own instincts …		

Table 2: Example of a pair of texts

3 Feature Selection

To determine the authorship of the two text chunks, we need to specify a text representation that can characterize the stylistic idiosyncrasies of each possible author. As a simple first solution, and knowing that only a small text size is available, we will focus on the most frequent word-types.

To generate a text representation, a tokenization must be defined. In this study, a token is a sequence of letters delimited by space or punctuation symbols. We also consider as token the punctuation marks (or sequence of them) such as comma, full stop, or question marks. Words appearing the `nltk` stopword list are included in this representation (179 entries composed of pronouns, determiners, auxiliary verb forms, some conjunctions and prepositions). Thus our strategy is based on word-types used recurrently by one author but not frequent in the underlying language (English in this case). One can compare this solution to the Zeta model [10], [11].

Then to determine the most frequent ones, the occurrence frequency (or term frequency denoted *tf*) of each word-type inside a chunk of text is computed. However, as each text chunk has a different size and we opt for a relative term frequency (*rtf*) computed as the term frequency divided by the text size. One can interpret this *rtf* value as an estimation of the probability of occurrence of the term in the underlying snippet. Finally, each pair is represented by the *rtf* of the k most frequent word-types, with k varying from 100 to 400.

4 Similarity Measure

Based on a vector of *k* elements reflecting the *rtf* of each selected word-type, the similarity (or distance) between the two text excerpts can be computed. In this study, we opt for the Labbé similarity [12]. This measure is normalized and returns a value between zero (nothing in common) and one (vectors are identical). All pairs of snippets with similarity above a given threshold (denoted δ) are considered to be authored by the same person. On the other hand, a similarity value lower than the specified threshold indicates different authors.

Denoting by d_1 and d_2 two document vectors, the Labbé distance corresponds to the ratio of the absolute difference of all n terms to the to the maximal distance between the two text representations as shown in equation 1.

$$\text{Dist Labbé}\,(d_1, d_2) = \frac{\sum_{i=1}^{n} \widehat{rtf}_{i,1} - rtf_{i,2}}{2 * n_{d2}} \; with \; \widehat{rtf}_{i,1} = rtf_{i,1} * \frac{n_{d2}}{n_{d1}} \qquad (1)$$

The decision rule is based on the value of Labbé similarity, which is (1 - Dist Labbé) (with δ = 0.5):

$$\text{Decision} \begin{cases} \text{Same author} & \text{if Sim Labbé } (d_1, d_2) > 0.5 \\ \text{Different authors} & \text{if Sim Labbé } (d_1, d_2) < 0.5 \\ \text{Non decision} & \text{otherwise} \end{cases} \qquad (2)$$

The implementation considers the absolute difference of all n terms in the text representations. For each term in the document, the difference between the absolute frequencies in Text d_1 and d_2 is computed. This requires both documents to have equal length. To ensure that both text have similar length, assuming Text d_1 (n_{d1}) is larger than Text d_2 (n_{d2}), we multiply the relative term frequency of Text $d_1(rtf_{i,1})$ with the ratio of the two lengths n_{d1} and n_{d2} as shown in Equation 1.

During the PAN CLEF 2020 author verification task, the system must return a value between 0.0 and 1.0 for each problem. In our case, the Labbé similarity score provide this value. In addition, we must specify "same-author", "different-authors" or provide a blank answer (meaning " I don't know") that will be considered as an unanswered question during the evaluation. Specify δ = 0.5 (see Equation 2), we ignore this last possible answer and we provide an answer to all problems.

5 Evaluation

As a performance measure, four evaluation indicators have been used. First, the AUC (area under the curve) is computed. This value corresponds to area under the curve generated according to the percentage of false positives (or false alarms) in the x-axis and the percentage of true positive cases in the y-axis over the entire test set. A model whose predictions are 100% wrong obtains an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Second the F_1 combines the precision and the recall into a unique value. In this computation, the non-answers are ignored. Third, c@1 is a variant of the conventional F_1 score which rewards systems leaving difficult problems unanswered. It takes into account both the number of correct

answers and the number of problems left unsolved [13]. Four, F_0.5_u is a measure according more emphasis when the system is able to solve the same-author cases correctly[14].

The entire system was based on only the training set, so the training and the evaluation was done directly on the same corpus. With 52,590 problems in the ground truth, the results of the similarity verification are shown in Table 3.

k	AUC	c@1	F_0.5_u	F_1	overall	Runtime
100	0.847	0.530	0.585	0.692	0.663	00:21:41.6
150	0.851	0.535	0.585	0.692	0.665	00:26:48.6
200	0.854	0.530	0.585	0.692	0.665	00:30:20.4
250	0.855	0.530	0.585	0.692	0.666	00:35:26.7
300	0.857	0.530	0.585	0.693	0.666	00:39:16.6
350	0.858	0.531	0.585	0.693	0.666	00:43:44.7
400	0.859	0.531	0.585	0.693	**0.667**	00:48:15.3
450	**0.860**	0.531	0.585	0.693	**0.667**	00:53:15.5
500	**0.860**	0.531	0.585	0.693	**0.667**	00:59:25.8

Table 3 : Evaluation based on different feature sizes

Increasing the number of features from 100 to 500 does not have a significant impact on the overall results as shown in Table 3. On the other hand, the run time is clearly increasing.

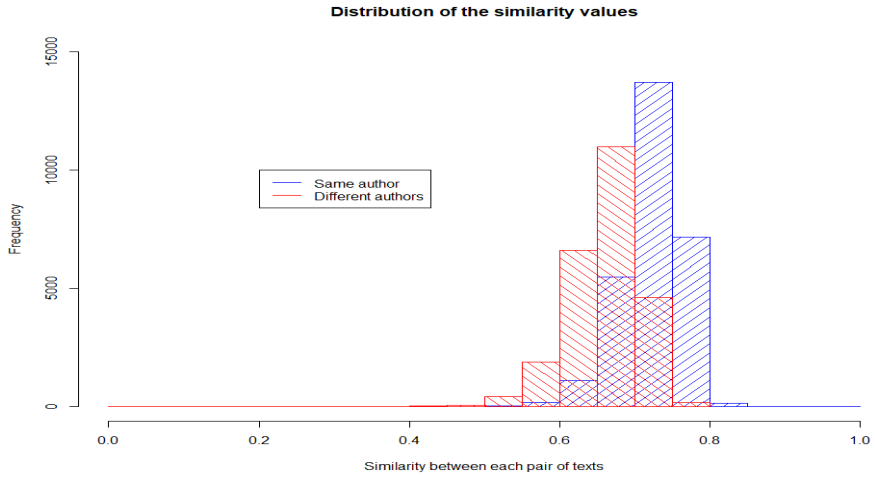

Figure 1 : Distribution of the similarity values for the two classes, same author or distinct authors (*k*=100)

To have a better view of the results, Figure 1 shows the distribution of the Labbé similarity values for the two classes, namely "same -author" and "different authors" (k=100). As one can see, the "same-author" distribution (in blue) presents a higher (mean: 0.723, sd: 0.041) and the distribution is more on the right (higher value) compared to the "different authors" distribution (mean: 0.660, sd: 0.048) and shown in red. However, the intersection between the two distribution is relatively large.

In a last experiment, instead of building the text representation with all possible word-types, we remove the 179 most frequent word appearing in the nltk stopword list. Table 4 reports the overall performance of both approaches with k=500. Depending on the evaluation measure, one representation strategy tends to propose the best effectiveness. The results are thus inconclusive.

	AUC	c@1	F_0.5_u	F_1	overall
With	**0.860**	0.531	0.585	0.693	0.667
Without	0.825	**0.591**	**0.618**	**0.718**	**0.688**

Table 4 : Evaluation with or without a stopword list (k=500)

Table 5 reports our official results achieved with the TIRA system [15] These evaluations correspond to our feature selection with the top 500 most frequent features including stop words.

AUC	c@1	F_0.5_u	F_1	overall
0.840	0.545	0.599	0.705	0.672

Table 5: Official Evaluation with (k=500)

6 Conclusion

Due to time constraint, this report proposes a simple text similarity technique to solve the authorship verification problem when facing with pairs of snippets. We proposed to select features by ranking them according to their frequency of occurrence in each text and taking only the most frequent ones (from 100 to 500) but in including the most frequent ones in the underlying language. With this proposed strategy, we want to identify terms occurring frequently by an author and frequent in the current language (English in this study).

The similarity computation is based on the Labbé between two vectors. The next step for us is to explore the reverse text representation, taking account only of the most frequent terms of a given language [16]. Of course, one could then combine the two results to hopefully improve the overall effectiveness.

References

[1] Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., Stein B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. *CLEF 2020 Labs and Workshops, Notebook Papers*

[2] Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, Martin Potthast: The Importance of Suppressing Domain Style in Authorship Analysis. CoRR abs/2005.14714 (2020)

[3] Stover, J., Winter, Y., Koppel, M., & Kestemont, M. (2015). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, *67*. https://doi.org/10.1002/asi.23460

[4] Suman, C., Saha, S., Bhattacharyya, P., & Chaudhari, R. (2020). Emoji Helps! A Multi-modal Siamese Architecture for Tweet User Verification. *Cognitive Computation*. https://doi.org/10.1007/s12559-020-09715-7

[5] Halteren, H. V. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing*, *4*(1), 1:1–1:17. https://doi.org/10.1145/1187415.1187416

[6] Iqbal, F., Khan, L. A., Fung, B. C. M., & Debbabi, M. (2010). E-mail authorship verification for forensic investigation. *Proceedings of the 2010 ACM Symposium on Applied Computing*, 1591–1598. https://doi.org/10.1145/1774088.1774428

[7] Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, *63*, 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

[8] Neal, T., Sundararajan, K., & Woodard, D. (2018). Exploiting Linguistic Style as a Cognitive Biometric for Continuous Verification. *2018 International Conference on Biometrics (ICB)*, 270–276. https://doi.org/10.1109/ICB2018.2018.00048

[9] Hirst, G., & Feng, V. W. (2012). Changes in Style in Authors with Alzheimer's Disease. *English Studies*, *93*(3), 357–370. https://doi.org/10.1080/0013838X.2012.668789

[10] Craig, H., & A.F. Kinney, A.F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

[11] Burrows, J.F. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1), 27-47.

[12] D. Labbé and C. Labbé. A tool for literary studies. Literary and Linguistic Computing, 21(3):311–326, 2006.

[13] Peñas, A., & Rodrigo, A. (2011). A Simple Measure to Assess Non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1415–1424. https://www.aclweb.org/anthology/P11-1142

[14] Bevendorff, J., Stein, B., Hagen, M., & Potthast, M. (2019). Generalizing Unmasking for Short Texts. *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)*, 654–659. https://doi.org/10.18653/v1/N19-1068

[15] Potthast, M., Gollub, T., Wiegmann, M., Stein, B. (2019) TIRA Integrated Research Architecture. In N. Ferro, C. Peters (eds), *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*. Springer, Berlin.

[16] Zhao, Y. & J. Zobel, J. (2007). Entropy-Based Authorship Search in Large Document Collection. In Proceedings ECIR2007, Springer LNCS #4425, 381-392.