A Model for Style Change Detection at a Glance

Notebook for PAN at CLEF 2018

Jamal Ahmad Khan

Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan

J_Ahmadkhan@Yahoo.com

Abstract. This year's PAN Author Identification sub-task for style change detection deals with a single question, whether or not a document has multiple authors? To answer this simple question, a simple straightforward and fast approach is proposed in this document. Some basic stylometry analysis techniques e.g. word frequencies (for stop-words and other POS words), punctuations, word pair frequencies and POS pair frequencies. In order to make fast comparison among word windows, a fast comparison model is built that can produce results in a glance. This model showed 65.1% accuracy over evaluation dataset and 63.83% accuracy over training dataset.

1 Introduction

Last year's CLEF PAN [1] challenge for "Style Change Detection", focused over finding the boundaries within the documents wherever a style change was detected. The proposed models [2, 3, 4] however showed low accuracies. So, this year's PAN [5] challenge is to simply detect whether or not a style change exists in a whole document. In order to answer that question, one must go through the full document with a handful of stylometry techniques. Documents in provided datasets may have zero to more number of textual chunks by different authors on same topic. Hence answering a simple question may take as much effort as finding the boundaries of style changes within a document. But this time one has to quit search for style changes wherever the detection model finds a change and mark the document as stylistically changed.

A model is presented that can detect stylometry changes in documents, in a way like some skilled human reader may normally detect in a glance [6]. Detailed methodology is explained in following sections.

1.1 Related Work

Let's have a review of previously used techniques for the task of intrinsic plagiarism detection. The authors [7] modeled a plagiarism detection method relying over text sentence features and outlier detection for stylistic feature changes. In another approach Bensalem et al [8] proposed a new text representation of whole documents using n-gram classes, where each n-gram class was based over least and most frequent words. In another approach the author [9] used a set of 36 text features to train her binary classifier for detection of plagiarized and non-plagiarized passages in documents.

2 Dataset

The training dataset of PAN at CLEF 2018 [5] for "Style Change Detection" includes a total of 2980 and a separate dataset for evaluation includes 1492 English documents over different topics. Both datasets included exactly half documents having a style change and half without any style change. However, the position and number of style changes for each documents was unknown. As the proposed model also takes into account the length of documents in terms of sentence counts, the following table shows documents lengths in both datasets.

Table 1. Number of Sentences in provided Datasets

Dataset	*< 25*	*≥ 25 - < 50*	*≥ 50 - < 60*	*> 60*
Training	593	1981	293	113
Evaluation	344	965	135	48

3 System Methodology

In order to detect style changes within a given set of documents in a shorter time period, only a subset of stylometric features was chosen. Also a "*divide and conquer*" strategy for quick processing of each document *D* was adapted. According to this document processing strategy, full text of each document will be divided into two or more sections. Each section will be processed independently from others, and in the end every two divided text sections will be compared for quick results.

Following are the processing steps for each document.

1. Text segmentation into sentences
2. Division of sentences into two or more groups

- Stylometric analysis for each group

3. Stylometric comparison and Style change calculation
4. Repetition of step 2 on basis of positive or negative results

3.1 Text Segmentation into Sentences

A text document in the dataset was segmented into sentences S_1,………. ,S_n; where each document *D* has *n* number of sentences. These sentences are assigned to an array *A*.

$$A = S_1, S_2, \ldots . S_i \,, S_{i+1} \ldots .., S_n \qquad (1)$$

$$i = \frac{n}{2} \qquad (2)$$

Where S_i is the middle sentence of array *A* index of each sentence and *n* is the number of total sentences in any document *D*. This array *A* is passed to a function *F*, that will perform following steps.

3.2 Division of Sentences into Two or More Groups

All sentences in array *A* are divided into two main sub-arrays A_j, A_k, as shown in following equations.

$$A_j = S_1, S_2, S_3 \ldots . S_i \qquad (3)$$

$$A_k = S_i, S_{i+2}, S_{i+3}, \ldots . S_n \qquad (4)$$

The sentence S_i is shared among both sub-arrays. Next steps involve separate word and character n-gram based stylometric analysis of both groups.

3.2.1 Favorite Stop Words

A list of fifteen most frequent English stop-words [10] was used to find the frequency of these words in each group.

$$FSW = \{ \text{the, of, and, a, to, in, is, it, that, you, for, have, I, not, on} \} \qquad (5)$$

Let FSW_j and FSW_kbe the favorite stop-words frequency lists. Following Table shows an example of *FSW* frequencies in both groups arranged in descending order.

Table 2. Arrangement of Favorite Stop-Words in Both Groups

Array	$\boldsymbol{StopWord_{frequency}}$									
FSW_j	of	a	the	and	to	you	I	not	………..	it
FSW_k	of	a	and	for	to	it	I	not	………..	is

3.2.2 Least Frequent Words

After removing stop-words [11] from both A_j, A_k, the Lf_j, Lf_k sets of 20 least frequent words are created respectively. The criteria for choosing a least frequent word simply depends on number of times it appears in each text group.

3.2.3 Most Frequent Words

After removing stop-words [11] from both A_j, A_k, a Mf_j, Mf_k set of 20 most frequent words are created respectively. The criteria for choosing a most frequent word simply depends on number of times it appears in each text group.

3.2.4 Most Frequent Word Pairs

A sliding window of size of two words and which moves ahead by single word was used to get Prf_j, Prf_k sets of 30 most frequent word pairs.

3.2.5 Punctuations

Punctuations appearing in both text groups A_j, A_k were arranged in descending order according to frequency of appearance as P_j, P_k respectively.

The number of stop-words, frequent words and word pairs was chosen and adjusted after several test runs of algorithm over test dataset. The motive of these adjustments was to figure out the least possible number of stylometric word n-grams proposed algorithm's speed and performance.

3.3 Stylometric Comparison and Style Change Calculation

Stylometric match score S among both sentence groups is calculated by using following formula, where each match among the members of stylometric sets will add to the final score.

$$\boldsymbol{S1} = \sum_{\boldsymbol{j,k=1}}^{\boldsymbol{n}} [P_j = P_k] + \sum_{\boldsymbol{j,k=1}}^{\boldsymbol{m}} [FSW_j = FSW_k] \qquad (6)$$

$$\boldsymbol{S2} = \sum \left(\left[Lf_j \cap Lf_k \right] + \left[Mf_j \cap Mf_k \right] + \left[Prf_j \cap Prf_k \right] \right) \qquad (7)$$

$$\boldsymbol{S} = \boldsymbol{S1} + \boldsymbol{S2} \qquad (8)$$

A recursive function *F* is used to carryout tasks like text stylometric comparison and style change detection on the basis of stylometric analysis. The functionality of *F* has been described above in the start of section 3.

The decision to recall *F* for A_j and A_k depends on following condition, where α is the threshold value for stylometric match and β is the least number of sentences that a document may contain for next function recall.

$$S \geq \alpha \, and \; n \geq \beta \qquad (9)$$

if n < β and $S \geq \alpha$, then *F* will return true, which means there is no style change in given document *D* and false otherwise.

Following figures will show the two function recalls of *F*. The arrows in following figures 1(b) and 2(d) shows the stylometric comparison of one text group with other.

Figure 1. (a) Shows Sentence Group Formation, (b) Shows Style Change Calculation

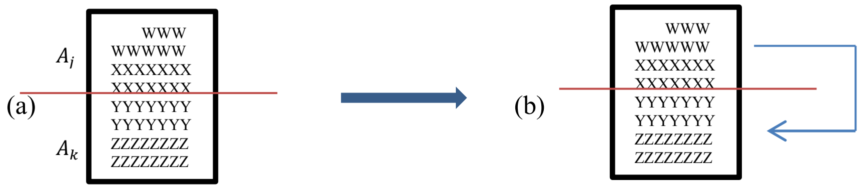
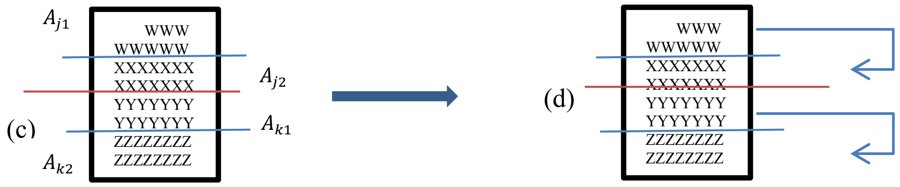


Figure 2. Showing *F* being recalled for both A_j and A_k separately and Style Change Calculation among A_{j1}, A_{j2} and A_{k1}, A_{k2} respectively



If any of the function's recall returns false, then it will be assumed that the document has a style change.

4 Results

Both α and β were adjusted after a number of experiments were carried out over Training dataset. After the threshold adjustments the model was ready to be tested over evaluation dataset. This model showed 65.1% accuracy over evaluation dataset and 63.83% accuracy over training dataset. The final score of proposed model is shown in following table.

Table 3. Results over Style Change detection Test Datasets

Accuracy	time
0.64275147929	00:01:10

The results show a consistent performance of model over all datasets. Also the model consumes least time from all other models presented in style change task.

5 Conclusion

The proposed model was built with one thing in mind, and that was to answer a simple question without carrying out complex and time consuming methodologies for style change analysis. This model achieved the first task in sense of least time consumption but however in terms of accuracy the results remained much lower than other presented techniques. This could however been improved by introducing more stylometric markers or via adding further recalls to function *F* for sub-groups of sentences.

Reference

1. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M.: Overview of the author identification task at PAN-2017: style breach detection and author clustering. In Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al. pp. 1-22 (2017).
2. Karas, D., Spiewak, M., & Sobecki, P.: OPI-JSA at CLEF 2017. Author Clustering and Style Breach Detection. Working Notes Papers of the CLEF (2017).
3. Khan, J. A.: Style Breach Detection: An Unsupervised Detection Model. Working Notes Papers of the CLEF (2017).

4. Safin, K., & Kuznetsova, R. : Style Breach Detection with Neural Sentence Embeddings. Working Notes Papers of the CLEF (2017).
5. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018).
6. Ehri, L. C.: Learning to read and spell words. Journal of Reading Behavior, 19(1), 5-31 (1987).
7. Kuznetsov, M., Motrenko, A., Kuznetsova, R., & Strijov, V. : Methods for Intrinsic Plagiarism Detection and Author Diarization. In: CLEF Working Notes, pp. 912-919 (2016).
8. Bensalem, I., Rosso, P., & Chikhi, S.: Intrinsic plagiarism detection using n-gram classes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1459-1464 (2014).
9. Rahman, R.: Information Theoretical and Statistical Features for Intrinsic Plagiarism Detection. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 144-148 (2015).
10. Most Common Words in English. Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Most_common_words_in_English (2018).
11. Doyle, D: Stopwords. Ranks NL, https://www.ranks.nl/stopwords (2018).