

Style Change Detection using Discourse Markers

Faisal Alvi^{1,2,*}, Hasan Algfri¹ and Naif Alqahtani¹

¹Information & Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

²Computer Science Program, Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan.

Abstract

This article presents our approach for the Style Change Detection Task at PAN 2022 using discourse markers. Discourse markers (such as ‘what’, ‘I have’, etc.) are words or expressions used to connect, organise and manage conversations. We present two different approaches for Style Change Detection at PAN-2022. For Task 1, (Style Change Basic) our approach is based on identifying conversational patterns within the documents between a user and a possible respondent. Then, using classification algorithms, we predict the point of style change detection within each document. For Task 2 (Style Change Advanced) and Task 3 (Style Change Real World), we use an extensive list of frequently occurring discourse markers to identify the number of speakers as the number of authors within the document. This prediction serves as the number of clusters for text segments within the document. Subsequently, using unsupervised clustering we detect clusters of similar text segments such that each cluster comprises of text segment groups corresponding to each author. The resulting F1 scores for our approaches on the test set are: 0.70518 for Task 1, 0.32128 for Task 2 and 0.56360 for Task 3.

Keywords

Style Change Detection, Discourse Markers, Conversational Patterns, Classification, Clustering

1. Introduction

The Style Change Detection [1] task has been available in PAN evaluation labs since 2017. For PAN-2022 [2], Style Change Detection consists of three tasks: (a) Style Change Basic, (b) Style Change Advanced, and (c) Style Change Real World. The datasets for all 3 tasks have been derived from user posts and replies on technology related issues. These posts simulate style change detection by providing a discourse between users and respondents on technology issues.

Discourse Markers [3] have been used for intrinsic plagiarism detection in the literature. Rao et al. [4] have used discourse markers as features towards intrinsic plagiarism detection on the PAN-2011 corpus. Likewise, Elamine et al. [5] have used discourse markers as features for intrinsic plagiarism detection on the PAN-16 and PAN-17 corpora. In this article we present our approach for style change detection using conversational patterns and discourse markers. Our approach utilises conversational patterns and frequently occurring unigram and bigram discourse markers in the first, second and third persons.

CLEF 2022: Conference and Labs of the Evaluation Forum 5–8 September 2022, Bologna, Italy


*Corresponding author.

✉ alvif@kfupm.edu.sa, alvi.faysal@gmail.com (F. Alvi); s201817820@kfupm.edu.sa (H. Algfri);

s201930770@kfupm.edu.sa (N. Alqahtani)

🆔 0000-0003-3827-7710 (F. Alvi); 0000-0002-8356-8683 (H. Algfri); 0000-0002-6343-4864 (N. Alqahtani)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

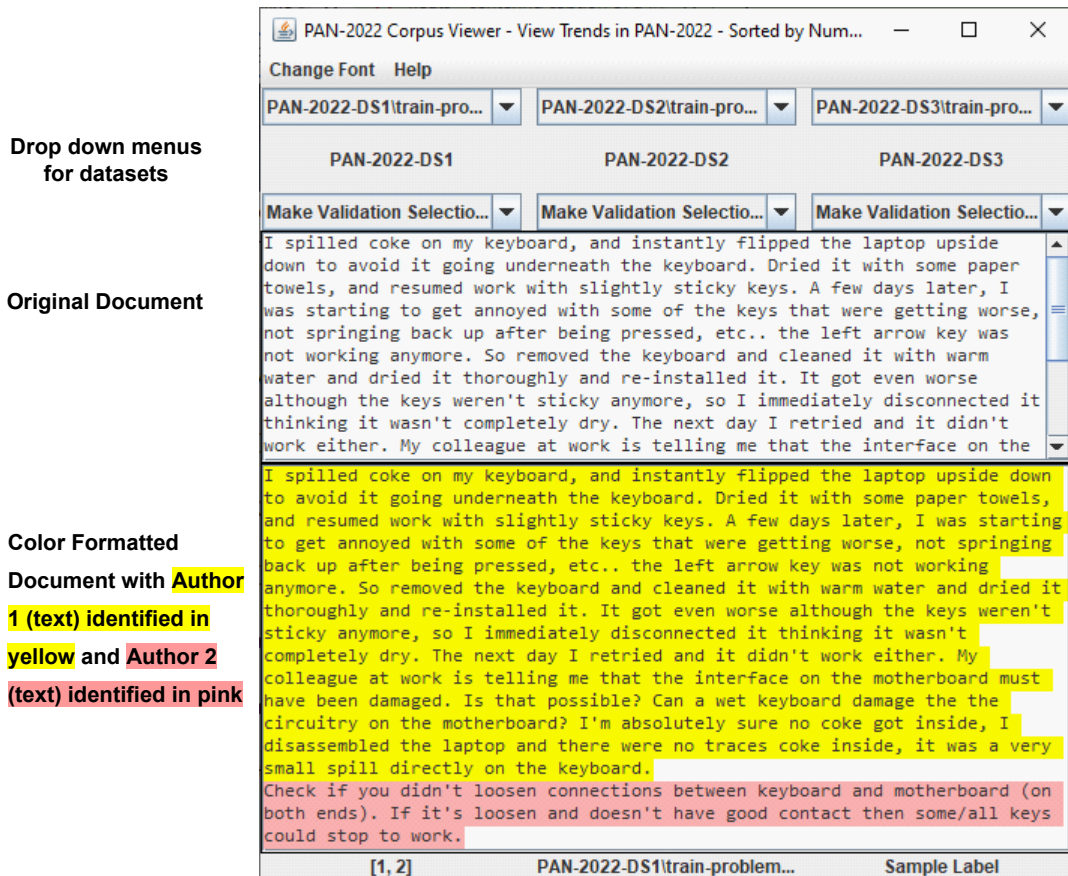


Figure 1: Corpus Viewer - a software tool to identify conversational patterns

2. Identifying Conversational Patterns using Corpus Viewer

Since the style change detection datasets are based on conversations around technology, we developed Corpus Viewer - a software tool to identify conversational patterns in Style Change Detection datasets, described as follows.

Corpus Viewer: A snapshot of Corpus Viewer is shown in Figure 1. This software tool consists of (a) drop down menus for selecting documents from datasets (train and validation), (b) a window displaying the original document, and (c) another window with color formatted text identifying various authors using the solutions provided for the training set.

Conversational Patterns: Using Corpus Viewer we identified a number of patterns that were present within the documents. For example, the file displayed in Figure 1 shows text highlighted by Author 1 with phrases such as: “*I spilled coke... so I immediately... I disassembled...*”. This text has a high presence of the first person pronoun ‘I’ in the first paragraph. In contrast, the text by Author 2: “*Check if you...*” has only a single instance of the second-person pronoun ‘you’ in the second paragraph. From this analysis we can conclude that this document represents an issue followed by an answer.

3. Proposed Approaches

In this section we describe the approaches employed for Task 1 and Tasks 2 & 3.

3.1. Task 1

We start by searching for patterns in dataset 1 observed using Corpus Viewer. More specifically, we search for words indicating conversational patterns (such as ‘Thank you’, ‘?’, verbs in their base form) at the boundary between Author 1 and Author 2. Based on a visual analysis of documents in dataset 1, we identified three different types of documents that were commonly found throughout the dataset as follows:

1. Question (Author 1) followed by an answer (Author 2), (e.g. the document in Figure 1),
2. A statement (Author 1) followed by a question (Author 2),
3. Two replies (Author 1 and Author 2) to an issue, which is not in the document.

Given a document, we attempt to identify whether it belongs to one of these types by searching for discourse markers. The rationale for identifying the document type is that each document type has some particular pattern, indicated by discourse markers at the point of change. For example, the presence of a verb in its base form at the beginning of a paragraph (such as ‘Check’, ‘Try’, ‘Do’) signifies an instruction by a respondent given to a user. Hence this probably identifies the point of Author change in a document of type 1. Likewise, the appearance of a polite word (such as ‘Thank you’ or ‘Thanks’) at the end of a paragraph identifies the point at which there is an Author change. Furthermore, appearance of a question mark at the end of a paragraph typically identifies a question.

Based on these patterns we construct feature vectors that incorporate all possible positions of author change for each document in a dataframe as shown in Table 1. For each row, appearance of identified patterns serves as the feature set for Author change. For example, the first row in Table 1 identifies the point of change of Authors between Paragraph 1 and Paragraphs 2, 3, 4.

Table 1

Dataframe Representing Point of Change between Authors for a given document

Author 1	Author 2	Pattern 1	Pattern 2 (...)	Author Change
Paragraph 1	Paragraphs 2, 3, 4	Yes	...	Yes
Paragraphs 1, 2	Paragraphs 3, 4	No	...	No
Paragraphs 1, 2, 3	Paragraph 4	No	...	No

After the dataframe construction, we apply four machine learning algorithms [6] on the training set (i.e., Decision Tree, Logistic Regression, Naive Bayes’ and Random Forest) as shown in Table 2, which represents the accuracy and F1 scores for each algorithm. These classification algorithms are used for predictions on the point of Author Change for each document in the validation set. From these values of accuracy and F1 Scores, we find that the Random Forest Classification Algorithm performs the best. Therefore we consider the prediction of the Random Forest Algorithm for a document, if available.

Table 2

Results of four Machine Learning Classification Algorithms (Bold values indicate the highest value)

	Decision Tree	Naive Bayes	Random Forest	Logistic Regression
Accuracy	0.80850	0.84306	0.89117	0.88089
F1 Score	0.62910	0.68144	0.70784	0.67106

3.2. Tasks 2 and 3

Tasks 2 and 3 require a more fine grained identification of multiple author changes. The number of authors for Tasks 2 and 3 ranges from 1 to 5 authors with change occurring possibly at each paragraph boundary. We use a two staged approach for identifying author changes for Tasks 2 and 3, stated as follows:

1. In the first step, we aim to identify the number of authors for each document. We consider the most frequent unigrams and bigrams in all the three datasets with the words 'I', 'we', 'you', 'he', 'she', 'they' and interrogative words as discourse markers. For each document, we construct a feature vector that includes counts of 212 first person, 185 second person, 41 third person and 10 interrogative unigrams and bigrams (such as "I'd", "you have", etc). In addition, statistical information such as number of paragraphs, number of words and characters, punctuation to text ratio and frequency of commas and question marks are also included as features at the document level. Subsequently, we use Random Forest Algorithm [6] (best performing, similar to task 1), to predict the number of authors in the test set for each document.
2. In the second step, using the number of identified authors predicted, we apply K-Means Clustering [6] to partition the paragraphs of a given document into n clusters, where n is the number of authors. K-Means Clustering is used as it is one of the most common algorithms used for unsupervised clustering. Features included for clustering of paragraphs are the same as that used in step 1, but at the level of paragraphs instead of at the document level.

Figure 2 illustrates this two-step process.

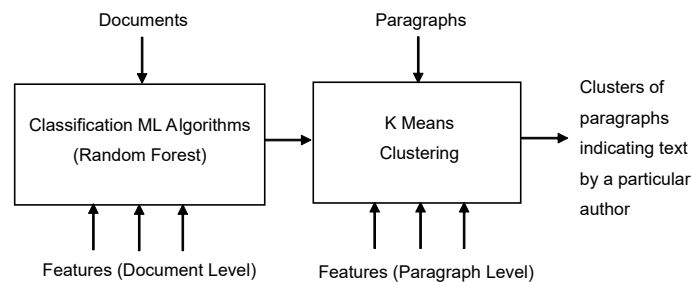


Figure 2: A two staged process for Style Change Detection for Tasks 2 and 3

The confusion matrices for the prediction of the number of authors for the validation datasets 2 and 3 in the first step are shown in Figure 3. It can be observed that the predictions are good for a single author as well as reasonable for documents with 5 authors. However, the model does not adequately discriminate between 2, 3 and 4 author documents for both datasets.



Figure 3: Confusion Matrices predicting the Number of Authors for Tasks 2 and 3

Predictions from this stage as the number of authors for each document are then sent as input to a clustering phase, where K-means clustering is used to cluster similar paragraphs into clusters, corresponding to the text by each author. The outcome of this stage are paragraphs with their corresponding authors as a solution.

4. Results and Discussion

Final runs on the test set were made on the TIRA platform [7]. The final results (F1 scores and Task 2 DER, JER) of the stated approaches on the validation and test sets are shown in Table 3.

Table 3

Results of F1 Scores on the Validation and Test Sets for Each Task (Dataset)

F1 Score	Task 1	Task 2	Task 3	Task 2 (DER)	Task 2 (JER)
Validation Set	0.70538	0.33016	0.57788	0.37989	0.52972
Test Set	0.70518	0.32128	0.56360	0.39240	0.52180

From these results we observe a very minor change from the F1 scores for the validation set to F1 scores for the test set. This suggests that there was no overfitting of the model. Furthermore, the approaches are effective in detecting style changes however these can be enhanced by further refining and improvement.

5. Conclusion

In this work, we presented two related approaches for detecting style changes for the Style Change Detection Task. Our approaches were based on finding conversational patterns as well as using discourse markers and were effective in detecting style changes in the documents. Future work involves developing these approaches by improved classification and clustering algorithms as well as addition of more features.

References

- [1] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, 2022.
- [2] 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: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. B. Cenedo, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.
- [3] B. Heine, G. Kaltenböck, T. Kuteva, H. Long, *On the Rise of Discourse Markers*, volume 219, John Benjamins Publishing Company, 2021.
- [4] S. Rao, P. Gupta, K. Singhal, P. Majumder, External & Intrinsic Plagiarism Detection: VSM & Discourse Markers based Approach—Notebook for PAN at CLEF 2011, in: V. Petras, P. Forner, P. Clough (Eds.), *Notebook Papers of CLEF 2011 Labs and Workshops*, 19-22 September, Amsterdam, The Netherlands, CEUR-WS.org, 2011. URL: <http://ceur-ws.org/Vol-1177>.
- [5] M. Elamine, S. Mechti, L. H. Belguith, Intrinsic Detection of Plagiarism based on Writing Style Grouping, in: *Language Processing and Knowledge Management*, 2017.
- [6] G. Bonaccorso, *Machine Learning Algorithms*, Packt Publishing Ltd, 2017.
- [7] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World, The Information Retrieval Series*, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.