Style Breach Detection: An Unsupervised Detection Model

Notebook for PAN at CLEF 2017

Jamal Ahmad Khan

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

J_Ahmadkhan@Yahoo.com

Abstract. This paper deals with the sub-task of PAN 2017 Author Identification, which is to detect style breaches for unknown number of authors within a single document in English. The presented model is an unsupervised approach that will detect style breaches and mark text boundaries on the basis of different stylistic features. This model will use some classical stylistic features like POS analysis and sentence lexical analysis. Also some new features naming common English word frequencies within sentence text, sentence expression and sentence attitude have been proposed. The new features may not be directly linked to author's style of writing but to the subject/topic of sentence under analysis. Moreover the model uses sentence window for style detection. The sentence window may be extended to neighboring sentences during its unsupervised analysis.

1 Introduction

Stylometry is an important tool in the field of digital text forensics, especially in cases where we have unidentified or dubious text documents [1] written by one or more authors. These documents do not have an external link, tool or repository to prove that which text passage relates to which author. In other words, we use stylometric approaches when we may have to ascertain if the acclaimed authorship of text document actually exists in circumstances where we do not have any external verification resources.

Stylometric approaches generally achieve higher accuracy for long documents [2] because longer documents contain more text to reveal stylistic features of authors like in the field of Intrinsic Plagiarism detection problem solving [3, 4]. But in cases of short documents or texts e.g. in cases of social media like twitter where there may be fewer sentences by each author, Stylometric approaches my not get more accurate results. Although much work has been done in cases of scam emails [5], cyber-crimes [6] and fake service provision reviews [7] using Stylometric models.

One way of using stylometric approach in case of author attribution and author profiling is by training the computer applications over specific writing style of some specific author in a number of documents. But as discussed above the task of

detecting style breaches within a document without knowing in advance about the exact number of authors is difficult task and also an objective for ongoing research. Detection of style breach is related to text segmentation where text boundaries are marked with detection in change of topics [8].

The presented model uses unsupervised classification approach to detect and mark passage boundaries in given documents on the basis of style breaches. A combination of well-known stylometric features like Syntactic, Lexical and content specific features [9] are used with features like ordinary words frequency, sentence expression and sentence attitude that may be related to textual topic specification and may not be directly related to author's style. But this approach may be very handy in cases where we want to relate one sentence to its neighboring sentences and thus detect exact passage boundaries within a given document.

Also this model is a good example of how a text as small as a sentence within a document may be helpful in finding its related sentences on the basis of stylometric and other parameters to help us figure-out the passage boundaries by unknown number of authors.

2 Dataset

The training dataset of PAN at CLEF 2017 [8] for the task of style breach detection under main task of author identification. The dataset contained about 187 English text documents of different lengths and sizes over different topics like biography, politics, travel, hotels etc. Along with each text document a truth file was provided which contained exact character positions indicating style breach occurrences within that document, topic of document however remains unchanged.

3 System Methodology

The presented model uses different types of classical stylometric methods along with some new methods in order to find text borders where style breach is identified. The system used sentences as text segmentation unit. The sentence window keeps extending over its neighboring sentences until style breach is detected. Following are the methodology steps used by the system in order to find out style breaches.

- Words lists preparation
- Text segmentation into sentences
- Sentence window based syntactic analysis
- Sentence window based lexical analysis
- Content based analysis of sentence window
  - Sentence window expression labeling
  - Sentence window attitude labeling
- Style breach calculation

3.1 Words Lists Preparation

Different types of lists of words were prepared from different internet sources [10, 11, 12, 13] that express specific moods or human feelings. Seven expression lists of words were used including anger, confusion, curiosity, urgency, satisfaction, inspiration and happiness; where all lists comprised of about 200 words each. One reason for choosing only these seven expressions was the availability of proper expressive words over internet sources for these expressions. The second reason was to use limited set of expressions that may express human feelings while writing some text. More expressions may be included for future research. Two words additional words lists of about 500 words each of which reflecting positive or negative attitudes [14, 15] were included. An example of these expressive and attitude lists is shown in table 1 and table 2.

Table 1. Example of words expressing different feelings

Index	*Expression*	*Words*
1	**Anger**	ordeal, outrageousness, provoke, repulsive ….
2	**Confusion**	doubtful, uncertain, indecisive , perplexed….
3	**Curiosity**	secret, confidential, controversial, underground…..
4	**Inspiration**	motivated, eager, keen, earnest….
5	**Happiness**	blissful, joyous, delighted, overjoyed…..
6	**Satisfaction**	accurate, satisfied, advantage, always…..
7	**Urgency**	magical, instantly, missing, quick……

Table 2. Example of words expressing positivity or negativity

Attitude	*Words*
Positive	admiring, adoring, affectionate, appreciative, approving….
Negative	abhorring, acerbic, ambiguous, ambivalent, angry, annoyed…...

An additional list of 5000 most common English words with word frequencies was also included [16] an example of which is shown in table 3. This list contributes in order to measure the commonality index in a sentence.

Table 3. Example of common English words with frequencies

Word	*Frequency*	*Word*	*Frequency*	*Word*	*Frequency*
A	10144200	casual	6946	Naval	4990
abandon	15323	casualty	6439	Near	54869
ability	51476	cat	21135	Nearby	13820
----------	----------	----------	----------	----------	----------

These lists became the part of model and will be used for labeling of sentences in next methodology steps.

3.2 Text Segmentation into Sentences

Each individual document *D* in the repository was segmented into sentences S_i, S_{i+1}, S_{i+2}, S_{i+3}, S_{i+4},…. S_n. A simple algorithm was used to break a document into array of sentences. It first traverse through each character of document *D* from start until the any of the two characters '.' or '?' are encountered, which indicates sentence endings. The sentence is extracted and the algorithm continues from next character as start of next sentence.

$$D = S_i + S_{i+1} + S_{i+2} + S_{i+3}, S_{i+4} + \ldots + S_n \quad (1)$$

Where *i* is the starting index of each sentence and *n* is the number of total sentences in *D*. The first three sentences of any document *D* will be the starting window W_j ($j = 1$) for initializing point that may or may not extend and merge with next adjacent sentence windows (two at a time) depending on further analysis, also the adjacent sentence windows W_{j+1} will also share boundary sentence as shown in equation 2 and 3.

$$W_j = S_i + S_{i+1} + S_{i+2} \quad (2)$$
$$W_{j+1} = S_{i+2} + S_{i+3} + S_{i+4} \quad (3)$$

The sentence S_{i+2} is common boundary sentence in first and second windows W_j and W_{j+1}. This common sentence among two adjacent windows will increase the similarity index when comparing both windows for a possible merger/extension.

As discussed above *n* is the total number of sentences in any document and each sentence window *W* can have only three sentences in start (as shown in equations 2 and 3); hence the maximum number of text windows in any document will be as shown in equation 4.

$$Max.\ Windows\ (m) = \frac{n}{3} \quad (4)$$

Let's consider for an example $j = 1$, so first two sentence windows W_1 and W_2 are chosen for further analysis. The next steps performed by model are as follows.

1. *Sentence Window based syntactic analysis:* Text in both adjacent windows is converted to its respective part of speech (POS) tags for each word present in texts as shown in table 4.

Table 4. Example of POS tagging in adjacent text windows

Window#	*Text*	*POS Tags*
W_1	Obama's mother returned to Hawaii in 1972 for five years, and then in 1977 went back to Indonesia, where she worked as an anthropological fieldworker. She stayed there most of the rest of her life, returning to Hawaii in 1994. She died of ovarian cancer in 1995.	NNP POS VBN TO NNP IN CD IN CD NNS, CC RB IN CD NN TO NNP, WRB PRP VBD IN DT JJ NN. PRP VBD RB JJS IN DT NN IN PRP$ NN, VBG NNS IN CD. PRP VBD IN JJ NN IN CD.
W_2	She died of ovarian cancer in 1995. Of his early childhood, Obama has recalled, "That my father looked nothing like the people around me that he was black as pitch, my mother white as milk barely registered in my mind." In his 1995 memoir, he described his struggles as a young adult to reconcile social perceptions of his multiracial heritage.	PRP VBD IN JJ NN IN CD. IN PRP$ JJ NN, NNP VBD, `` IN PRP$ NN VBD NN IN DT NNS IN IN PRP VBD JJ IN NN, PRP$ NN JJ IN NN VBN IN PRP$ NN. IN PRP$ CD NN, PRP VBD NNS IN DT JJ NN TO VB JJ NNS IN JJ NN.

From the two examples presented in table 4, the model extracts following text features:

Starting and ending POS tags (t_{1_j}, t_{2_j}) for each sentence in each sentence window e.g. starting POS tags for $W_1$are t_{1_1} = {NNP, PRP, PRP} and ending POS tags are t_{2_1} = {NN, CD, CD}.

Most frequent POS tags and POS tag pairs (t_{3_j}, t_{4_j}) are extracted e.g. most frequent POS tag in W_1 and W_2 is t_{3_1}, t_{3_2} = IN and most frequent POS tag pairs in both windows are t_{4_1} = {IN, CD} and t_{4_2} = {IN, PRP$} respectively.

2. *Sentence Window based Lexical Analysis:* At this step, the model performs a lexical analysis for both text windows. In this analysis following features are extracted:

Most frequent alphanumeric and non-space character (d_{1_j}) in the text window is extracted e.g. d_{1_1} = *'e'* in both text windows in shown table 4.

Most frequent non-alphanumeric and non-space character ($\boldsymbol{d_{2\_j}}$) in the text window is extracted e.g. $\boldsymbol{d_{2\_1}}$, . $\boldsymbol{d_{2\_2}}$ = *','* in both text windows W_1 and W_2 .

Most frequent word ($\boldsymbol{d_{3\_j}}$) in the text window is extracted where *i* in equation below is the index of word *w* e.g. $\boldsymbol{d_{3\_1}}$ = *"in"* and $\boldsymbol{d_{3\_2}}$ = *"of"* in both text windows respectively as mentioned in table 4. The frequency of each word w_i is calculated as shown in equation 5.

$$\textit{Word Frequency } (f_i) = \sum_{i=1}^{n} w_i \qquad (5)$$

Character to Space Ratio CR_j is calculated for each text window as shown in equation 6.

$$\textit{Character to Space Ratio } (CR_j) = \frac{Num.\ of\ spaces}{Num.\ of\ non{-}space\ chars.} \qquad (6)$$

3. *Content Based Analysis of Sentence Window*: At this step commonality index CI_j of each window is calculated using the list *L* of 5000 common words. Let w_i be a common word existing in both *L* and any text window W_j where *i* specifies the index (*i = 1… 5000)* in *L* in eq. 7.

$$Commonality\ Index\ (CI_j) = \sqrt[l]{(\sum_{k=1}^{n} cf_i \times lf_i)} \qquad (7)$$

Where *k* is the total number of coexisting words in both *L* and W_j, and cf_i be the frequency of w_i in W_j, lf_i is the frequency of w_i in list *L* (as shown in table 3) and *l* is the total number of words in W_j.

Next two steps can be considered as sub-steps of Content based analysis.

4. *Sentence Window Expression Labeling:* The model will label each window with a specific feeling or human mood expression e_j. Let *i* is the index (*i = 1… 7)* of expression list E_i as shown in table 1, Let w_m be a coexisting word in both E_i and text window W_j where *m* specifies the index in W_j. Expression score e_i is measured on the basis of following equation.

$$Expression\ Score\ (\boldsymbol{e_i}) = \sum_{k=1}^{n} ef_i \qquad (8)$$

Where *k* is the total number of coexisting words in both E_i and W_j, and ef_i be the frequency of w_m in W_j. After calculating all seven expression scores the model will calculate *e* through following equation.

$$e_j = \max_{i=\{1,\ldots,7\}} e_i \qquad (9)$$

In cases where two or more expression scores are equal, or all expression scores are zero, the model will assign a "*neutral"* expression for window W_j.

5. *Sentence Window Attitude Labeling*: The model will label each window with a specific attitude or human behavior a_j. Let *i* is the index (*i = 1… 2)* of attitude list A_i as shown in table 2, Let w_m be a coexisting word in both A_i and text window W_j where *m* specifies the index in W_j. Attitude score a_i is measured on the basis of following equation.

$$Attitude\ Score\ (\boldsymbol{a_i}) = \sum_{k=1}^{n} af_i \qquad (10)$$

Where *k* is the total number of coexisting words in both A_i and W_j, and af_i be the frequency of w_m in W_j. After calculating both positive and negative attitude scores the model will calculate *a* through following equation.

$$a_j = \max_{i=\{1,\ldots,2\}} a_i \qquad (11)$$

In case both scores are equal or zero, the model will assign a neutral attitude for W_j e.g. both $W_1$and W_2 have neutral attitude.

6. *Style Breach Calculation*: After computing above mentioned stylistic and other attributes we get two result sets naming V_1, V_2 and two matrices v_1 and v_2 for text windows W_j and W_{j+1} respectively

$$V_1 = \{t_{1\_1},\ t_{2\_1},\ t_{1\_1},\ t_{2\_1},\ d_{1\_1},\ d_{2\_1},\ d_{3\_1},\ e_{1,}\ a_1\} \qquad (12)$$

$$V_2 = \{t_{1\_2},\ t_{2\_2},\ t_{1\_2},\ t_{2\_2},\ d_{1\_2},\ d_{2\_2},\ d_{3\_2},\ e_{2,}\ a_2\} \qquad (13)$$

$$v_1 = [\ CR_1 \quad CI_1\ ] \qquad (14)$$

$$v_2 = [CR_2 \quad CI_2\ ] \qquad (15)$$

The system will now measure stylistic similarity score α as shown in following equations

$$V_1 \cap V_2 = \{x:\ x \in V_1\ and\ x \in V_2\} \qquad (16)$$

Where, for each x in equation 15, the similarity score α is incremented accordingly. v_1 and v_2 are treated separately as matrices because these two contains decimal values. A matrix subtraction is applied to v_1 and v_2

$$abs\ (v_1 - v_2\ ) = [cr \quad ci] \qquad (17)$$

If *cr* and *ci* lie within a threshold range θ_1 described in next section, then similarity score α is incremented accordingly. Finally, it's time to decide whether or not to merge W_j and W_{j+1} on the basis of value of α lies within a threshold range $\theta_2$described in next section. At this point two cases will emerge:

Case-1: **α lies within a threshold range θ_2**

In this case W_j and W_{j+1} are considered merged, and a new resultant window W_r will be created where *r* is the index of resultant window. The model will continue from step 1 of methodology for sentence W_{j+1} and W_{j+2}.

$$W_r = S_i + S_{i+1} + S_{i+2} + S_{i+3} + S_{i+4} \quad (18)$$

W_r will keep expanding until *case-1* keeps occurring and this resultant window will reflect a single style for all sentences contained within.

Case-2: **α does not lie within a threshold range θ_2**

In this case the coexisting sentence in both adjacent windows will stay either in window W_j or in W_{j+1} e.g. let's assume S_{i+2} in equations 2 and 3.

1. S_{i+2} will become a separate single sentence window W_c.
2. Stylistic score is calculated for W_c following same methodology steps and its distance from both W_j and $W_{j+1}$is calculated.
3. S_{i+2} may remain in either of the two sentence windows depending on the distance value calculated.
4. If S_{i+2} remains in W_j then W_{j+1}will be restructured for next consecutive sentences as shown below.

$$W_{j+1} = S_{i+3} + S_{i+4} + S_{i+5} \quad (19)$$

5. If S_{i+2} remains in W_{j+1} then W_j will be restructured as shown below.

$$W_j = S_i + S_{i+1} \quad (20)$$

After the style breach detection among first two consecutive sentence windows, new windows W_{j+1} and W_{j+2}will be compared starting from step 1 of methodology.

In the end we have a set of resultant windows known as $R = W_{r=\{1....m\}}$ where *m* is the maximum number of sentence windows and each W_r in *R* is considered a breach detection.

4 Results

A number of experiments were carried out in order to adjust the threshold values $\theta_1$and θ_2 for which the final F-Measure score was highest. Once the values were adjusted over the training dataset, the system was ready to run for test dataset provided at TIRA [17] in order to detect style breaches.

Following are the evaluator results shown in table 5.

Table 5. Training and Test Results over Style Breach detection Datasets

Corpus	*Win. Diff*	*Win. Precision*	*Win. Recall*	*Win.F-Measure*
Training dataset	0.5184	0.3656	0.4841	0.2671
Test dataset	0.4799	0.39900	0.48710	0.2888

The results were improved for the final test dataset, however the model precision remained low from recall and that affected the final F-Measure score, which shows that more experiments over different data sources for adjusting threshold values may be required.

4 Conclusion

In this paper an unsupervised model for the detection of style breach is presented, this research field is rather new and more difficult to implement because non availability of any external resources for reference and also we only have to rely on stylistic attributes of unknown number authors that may or may not have contributed in the creation of text document under inquiry, hence this model presents new directions or ways i.e. Expression and Attitude labeling of textual windows in order to find style breach within sentences without the pre-assumption of authors style of writing and relying more on text content. In future the results can be improved with discovery of more text labels or with the addition of more expression lists and reduction of conventional stylistic approaches, this model can hence be applied to other languages as well.

References

1. [Online] https://en.wikipedia.org/wiki/Stylometry, (2017)
2. Brocardo Marcelo Luiz, Issa Traore, Sherif Saad. Authorship verification for short messages Using stylometry. Computer, Information and Telecommunication Systems (CITS), International Conference (2013)
3. Benno Stein, Barrón Cedeño, Eiselt, Martin Potthast, Paolo Rosso. Overview of the 3rd international competition on plagiarism detection. In: CEUR Workshop Proceedings. CEUR Workshop Proceedings (2011)
4. Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim Strijov. Methods for Intrinsic Plagiarism Detection and Author Diarization Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, Évora, Portugal,. CEUR-WS.org. ISSN 1613-0073 (2016)
5. Edoardo Airoldi, Bradley Malin. Data mining challenges for electronic safety. The case of fraudulent intent detection in e-mails. In Proceedings of the Workshop on Privacy and Security Aspects of Data Mining (2004)
6. B. Sullivan. Seduced into scams: Online lovers often duped. MSNBC (2005)
7. Audun Josanga, Roslan Ismailb and Colin Boyda. A survey of trust and reputation systems for online service provision. Decis. Support Syst. 43, 2, 618–644 (2007)
8. Michael Tschuggnall, Efstathios Stamatatos, Ben Verhoeven, Walter Daelemans, Gunther Specht, Benno Stein and Martin Potthast. Identification Task at PAN 2017: Style Breach Detection and Author Clustering. In: (Eds.) CLEF Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 10456 (2017)
9. Ahmed Abbasi, Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), Volume 26 Issue 2, Article No. 7 (2008)
10. [Online] http://www.manythings.org/vocabulary/lists/l (2017)
11. [Online] https://www.vocabulary.com/lists/202236 (2017)
12. [Online] http://descriptivewords.org/descriptive-words-for-attitude-personality (2017)
13. [Online] http://www.english-at-home.com/vocabulary/words-that-describe-behaviour (2017)
14. [Online] http://positivewordsresearch.com/list-of-positive-words (2017)
15. [Online] http://www.enchantedlearning.com/wordlist/negativewords.shtml (2017)
16. [Online] http://www.wordfrequency.info/free.asp?s=y (2017)
17. [Online] http://www.tira.io/tasks/pan/ (2017)