BLN-Gram-TF-ITF as a Language Independent Feature for Authorship Identification and Paragraph Similarity

Nawaf Ali
Al-Isra University
Jordan
nawaf.ali@iu.edu.jo

Roman Yampolskiy
University of Louisville
United States
roman.yampolskiy@louisville.edu

***Abstract*-Authors tend to leave traits in their writings, even if they try not to. By analyzing these traits by looking for textual features, one can construct a form of authorial profile that can distinguish ones writing from another; this is known as Authorship identification. The BLN-Gram-TF-ITF has been implemented as a new feature to identify authors by analyzing samples of their writings. New experiments demonstrated that BLN-Gram-TF-ITF feature is also a language independent, and can be used to measure paragraphs similarities within a book or between different books.**

Keywords—Authorship; TFIDF; N-grams; Stylometry; Text features; Text similarity.

1. INTRODUCTION

Authorship attribution is a behavioral biometric, in which one can identify the author of a document by extracting textual features and map it to the potential author of that text. Yampolskiy et al conducted a survey on different behavioral biometrics [1]. Different textual features are present and are being used in text mining and classifications. Stamatatos has a comprehensive survey on modern Authorship attribution methods and text features used in those methods [2].

Text similarity research started in the early 1970s in Information Retrieval (IR). Similarity measured between retrieved text and original text in database was used in this regard[3]. Nowadays, text similarity gained great importance in Natural Language Processing (NLP), text classification, web retrieval and question answering system[4].

Text similarity algorithms use the vector space and cosine measure as the way to represent text similarities. This ignores semantic similarities between terms. In order to solve this problem, Huang et al used paragraph random walk algorithm [4].

Liu et al proposed a dynamic multi-document summarization algorithm using the Text Similarity Computing Method (TSCM)[3]. Dynamic multi-document summarization is very useful in News Information Detection (NID)[5].

2. AUTHORSHIP ATTRIBUTION CATEGORIES

Authorship attribution can be categorized as followed:

- Authorship Identification: Given an unclaimed document(s), one can identify the correct author from the available potential authors.
-
- Authorship Verification: Given a certain text, one can verify if a certain author did actually write the text or not.
- Plagiarism Detection: Identifying copied materials from one source to another without being referenced. This is a good example of paragraph similarity application.
- Author Profiling: extracting authorship information and constructing a profile for ones writings.
- Detection of stylistic inconsistencies: Detection of multiple traits if we have more than one author writing a document [6].

2.1. TEXT FEATURES

2.1.1. TFIDF

In information retrieval domain, vector space model elements represent corpus documents. After tokenizing and stemming these documents, Euclidean space axes represent each token. These documents are vectors in this n-dimensional space. For each token (term) in a document (d)with (N) documents,one can define the Inverse Document Frequency (IDF) as:

$$\mathrm{IDF} = 1 + \log\left(\frac{n_t}{N}\right) \quad (1)$$

The term (n_t / N) in equation 1[7]represents the rarity of the term (t), such that (n_t) is number of documents having the term (t) in them over number of all documents(N). This rarity measure is also considered as an importance score for that term.

Term Frequency (TF) is another measure for calculating the number of times term (t) occurs in document (d) relative to the total number of terms in that document as shown in equation 2.

$$\mathrm{TF} = \frac{\mathrm{freq}(t,d_t)}{\|d_t\|} \quad (2)$$

Such that freq(t,d_t) will calculate the frequency of term (t) in document (d_t), and the $\|d_t\|$ is the number of terms (tokens) in document (d_t)[7].

So the TFIDF will be:

TFIDF = $TF \times IDF$ (3)

Another form of the TFIDF will have different IDF calculation, as shown below in equation 4.

$$\text{IDF} = \log\left(\frac{\|N\|}{1+n_t}\right) \quad (4)$$

The one is added to the denominator to avoid division by zero if the term frequency for that term in that document is zero.

What does IDF represent?Assuminga term (t) appears in all the documents in the corpus, this will lead to the values of ||N|| and (n_t) to be the same, and the log will be zero, and the IDF will equal 1 from equation 1, and the TFIDF will equal to TF in this case, and a value close to zero in equation 4[6].

The closer the TFIDF value gets to zero, the less weight this term have to classify that document[8].

2.1.2. N-GRAMS

When dealing with N-Grams, there are two levels we are interested in: Token (word) level grams, and the byte (character) level grams.

A. TOKEN N-GRAMS

Token N-Gram is one of the first text features used in Stylometry. In this model, one selects N successive words or tokens from the text, as if onehas a sliding window of size N moving over the text. In all cases, N is a positive number, and will determine the sliding window size. When N=1, the resulting gram will be what is known as a Bag Of Words (BOW), bigrams for N=2, and trigrams for N=3 as seen in Fig 1.

Different studies have shown different preferences for choosing N. In general, best results for authorship attribution were achieved for values of N>=3[9, 10].

"I played with my brother this morning"

I played with	played with my	with my brother	my brother this	brother this morning

Figure 1: Token- based N-Gram Example for N=3[6]

B. BYTE LEVEL N-GRAMS

Byte Level N-Grams are a very useful feature for studying the style of authors and also used for deciphering messages.Looking for possible bigrams and trigrams and other ngrams of characters, and from the frequencies of those ngrams, one can decipher the text and get the original message.

Stamatatos et al., used the byte level N-grams to identify the developer of a source code [11], he also used the same technique to detect plagiarism [12].

Generally speaking, N-grams, whethertoken or byte level, did outperformed other features when used for authorship attribution.

"I played with you"

'I p'	' pl'	'pla'	'lay'	'aye'	'yed'	' wi'	'wit'	'ith'	'th '	'h y'	' yo'	'you'

Figure 2: Character N-Gram for N=3[6]

C. BAG OF WORDS (BOW) FEATURE

As it was mentioned earlier, the BOW is a special case of the token ngrams when N=1, by this, one will count the frequency of words in a document, and there are two possible ways to do the feature in this case, either a Boolean value if the word is present or not, or an integer value of how many times did a word appear in the document.

3. BYTE LEVEL N-GRAM TERM FREQUENCY INVERSE TOKEN FREQUENCY (BLN-GRAM-TF-ITF)

Ali et al proposed this feature [6]. The idea behind this feature came from observing the importance of the N-Gram feature and the TFIDF in classification.Several research studies showed increased accuracy when using either one of these two features [8, 9, 13-16]. Treating the text as characters rather than tokens as explained in Fig. 2, the Byte-Level N-Gram slides over the characters and forms the Tri-Grams whenthe value of N =3[6].

BLN-Gram-TF-ITF will implement the idea of TFIDF but with different perspective.In this case, the token will be dealt with as TFIDF deals with documentsand the terms will be dealt with as TFIDF deals with words,the terms in this case will be the trigram generated from the Tri-Gram List.

For each unique term in Tri-Gram list, the TFITF will be calculated as followed:

$$TF = \frac{freq(t,Ngram)}{\|Ngram\|} \quad (5)$$

Such that, the calculated frequency will be based on the times this term generated from the Tri-Gram occurred in the Tri-Gram list divided over the length of Tri-Gram list. The ITF will be calculated as follows:

$$ITF = \log\left(\frac{\|NT\|}{n_{T\ with\ t}}\right) \quad (6)$$

Where ||NT|| is the total number of tokens in the corpus, and the ($n_{T\ with\ t}$) is how many tokens (T) containing the term (t), which corresponds to the IDF frequency of documents having token T. Fig.3 presents the flow chart for BLN-Gram-TF-ITF after combining with BOW.

4. Experiments and Results

Experiments were conducted on two sets of corpora, one German books from Gutenberg project [17]. Second experiments were using paragraphs from English books from Gutenberg project [17].

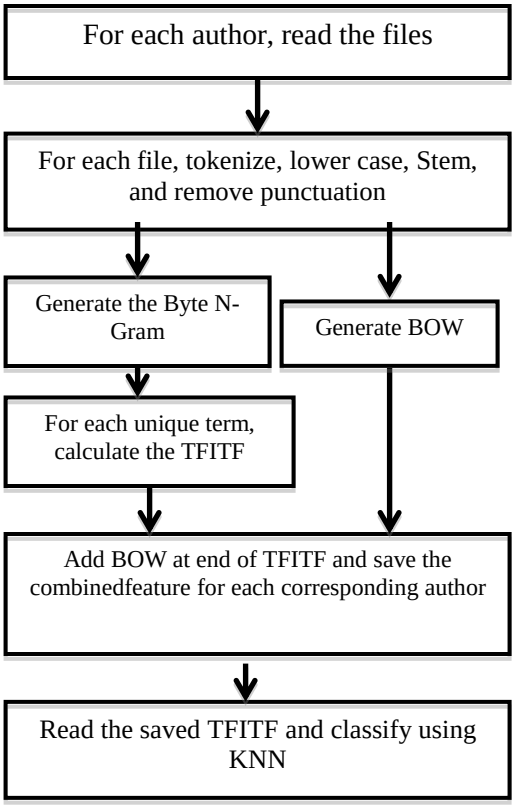

Figure 3: The Byte Level N-Gram Term Frequency Inverse Token Frequency Combined with BOW Flow Chart

a. GermanBooks Experiments

Ten different books written in German language were used, 10 files with 500 words each per book. The following are the books used in this experiment:

- Faust, Eine Tragödie by Johann Wolfgang von Goethe.
- Phantasten by Erich von Mendelssohn Release.
- Durch Wüste und Harem Gesammelte Reiseromane, Band I by Karl May.
- Der Untertan by Heinrich Mann.
- Buddenbrooks Verfall einer Familie by Thomas Mann.
- Das rasende Leben Zwei Novellen by Kasimir Edschmid.
- Die sechs Mündungen Novellen by Kasimir Edschmid.
- Die Fürstinby Kasimir Edschmid.
- Timur Novellen by Kasimir Edschmid.
- Über den Expressionismus in der Literatur und die neueDichtung by Kasimir Edschmid.

This experiment was performed just to demonstrate if the BLN-Gram-TF-ITF is language independent or not, it was tested to work on English language by Ali et al [6]. The experiments on the German corpus yield an average accuracy of 85%. This demonstrates that the BLN-Gram-TF-ITF is language independent.

b. Paragraphs from English Books Corpus Experiments

Six Authors were selected for the Experiments, and 10 paragraphs per author were used with 500 words per paragraph. The corpus used the following books:

1. Bleak House by Charles Dickens
2. Mansfield Park by Jane Austen.
3. The Adventures of Tom Sawyer by Mark Twain
4. The Parent's Assistant by Maria Edgeworth.
5. Moby Dick by Herman Melville.
6. Hamlet by William Shakespeare.

One sample paragraph was tested from each book with 500 words each.The results are shown in Table 1.

Table 1: Similarities between paragraphs from books as listed above

The results interestingly showing a higher similarity between the paragraphs chosen from Melville and Shakespeare, otherwise the similarities between all other paragraphs are low.

Book #	Paragraph from Book #					
	1	2	3	4	5	6
1	1	0.073	0.1192	0.1708	0.1877	0.1025
2	0.073	1	0.0657	0.2487	0.0737	0.0321
3	0.1192	0.0657	1	0.0872	0.1136	0.0484
4	0.1708	0.2487	0.0872	1	0.1082	0.0727
5	0.1877	0.0737	0.1136	0.1082	1	0.5856
6	0.1025	0.0321	0.0484	0.0727	0.5856	1

c. Experiments with Combining BOW with BLN-Gram-TF-ITF

As seen in Fig. 3, the BOW feature was combined with the BLN-Gram-TF-ITF by assigning the BOW values to the end of the BLN-Gram-TF-ITF feature. Fig. 4 shows that average accuracy did change dramatically for small files, and was fluctuating for larger files.

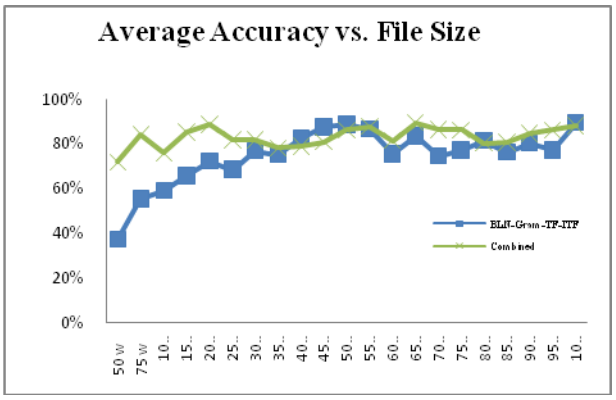

Figure 4: Average accuracy vs. file size between BLN-Gram-TF-ITF and same feature combined with BOW.

Performing paired t-test to test for statistical significance between the accuracy with and without BOW combined, Table 2 shows that the values obtained after combining the BOW feature with the BLN-Gram-TF-ITF is actually statistically significant with P-Value of 0.002.

Table 2: Paired t-test for Statistical significance between the two average accuracies of BLN-Gram-TF-ITF and Combined Feature. 95% CI for mean difference: (-0.1314, -0.0362).

	N	Mean	StDev	SE Mean
C1	21	0.7469	0.1241	0.0271
C2	21	0.8306	0.0456	0.0099
Difference	21	-0.0838	0.1046	0.0228

T-Test of mean difference = 0 (vs not = 0):
T-Value = -3.67 P-Value = 0.002

4. Conclusions and Future Work

The BLN-Gram-TF-ITF was tested to identify the original author of a text for both English text and German text. BLN-Gram-TF-ITF showed that it is a language independent feature; an average accuracy of 85% was achieved for the German language when experimented to identify the authors of 10 different books.

Combining the BOW feature with the BLN-Gram-TF-ITF did show an increase average accuracy especially for small files.

The Feature was created this time with applying stem porter. Removing stop words when testing for similarities and keeping the stop words when looking for authorial traits.

The BLN-Gram-TF-ITF feature was experimented with cosine similarity and did show a sound results. The experiments were tested on paragraphs from within the same book and paragraphs from different books for different authors.

More features need to be tested and combined with the BLN-Gram-TF-ITF and test for accuracy. In addition, further experiments will be conducted on measuring similarities for foreign languages and wider range of books for different authors.

References

[1] R. V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *International Journal of Biometrics (IJBM).* vol. 1, pp. 81-113, 2008.

[2] E. Stamatatos, "Author Identification Using Imbalanced and Limited Training Texts," pp. 237-241, 2007.

[3] M. Liu, D. Zheng, T. Zhao, Y. Yu, and J. Zhou, "Text similarity cumulative model and algorithm research for dynamic multi-document summarization," *Journal of Computational Information Systems,* vol. 7, pp. 1698-1705, 2011.

[4] X. Huang, J. Zhang, H. Chen, and W. Chen, "Research on Text Similarity Algorithm Based on Paragraph Random Walk⋆," *Journal of Computational Information Systems,* vol. 9, pp. 9103-9110, 2013.

[5] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, *et al.*, "Topic-based novelty detection," in *1999 Summer Workshop at CLSP Final Report. Available at* http://www. *clsp. jhu. edu/ws99/tdt*, 1999.

[6] N. Ali and R. V. Yampolskiy, "BLN-Gram-TF-ITF as a new Feature for Authorship Identification," in *The Third ASE International Conference on Cyber Security*, Stanford, CA, USA, 2014 (Under Review).

[7] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*: Wiley-IEEE Press, 2011.

[8] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[9] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, 2003, pp. 255-264.

[10] P. Juola, "Authorship attribution," *Foundations and Trends in information Retrieval,* vol. 1, pp. 233-334, 2006.

[11] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski, and B. S. Howald, "Identifying authorship by byte-level n-grams: The source code author profile (scap) method," *International Journal of Digital Evidence,* vol. 6, pp. 1-18, 2007.

[12] E. Stamatatos, "Intrinsic plagiarism detection using character n-gram profiles," *threshold,* vol. 2, pp. 1,500, 2009.

[13] A. Mohan, I. M. Baggili, and M. K. Rogers, "Authorship attribution of SMS messages using an N-grams approach," CERIAS Tech Report 20112010.

[14] J. Houvardas and E. Stamatatos, "N-Gram Feature Selection for Authorship Identification," in *Artificial Intelligence: Methodology, Systems, and Applications*. vol. 4183, J. Euzenat and J. Domingue, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 77-86.

[15] S. Gianvecchio, X. Mengjun, W. Zhenyu, and W. Haining, "Humans and Bots in Internet Chat: Measurement, Analysis, and Automated Classification," *Networking, IEEE/ACM Transactions on*, vol. 19, pp. 1557-1571, 2011.

[16] T. Matsuura and Y. Kanada, "Extraction of authors' characteristics from Japanese modern sentences via n-gram distribution," in *Discovery Science*, 2000, pp. 315-319.

[17] Gutenberg. (2012, Dec, 20). *Project Gutenberg*. Available: http://www.gutenberg.org