Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts

Helena Gómez-Adorno[1,2], Juan-Pablo Posadas-Duran[3], Germán Ríos-Toledo[4],
Grigori Sidorov[1], Gerardo Sierra[2]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación,
Ciudad de México, Mexico

[2] Universidad Nacional Autónoma de México, Instituto de Ingeniería,
Ciudad de México, Mexico

[3] Instituto Politécnico Nacional (IPN), Escuela Superior de Ingeniería Mecánica y Eléctrica
Unidad Zacatenco (ESIME-Zacatenco), Ciudad de México,
Mexico

[4] Centro Nacional de Investigación y Desarrollo Tecnológico,
Cuernavaca, Mexico

hgomeza@iingen.unam.mx, german_rios@cenidet.edu.mx,
gsierram@iingen.unam.mx, jposadasd@ipn.mx, sidorov@cic.ipn.mx

Abstract. In this paper, we present an approach to identify changes in the writing style of 7 authors of novels written in English. We defined 3 stages of writing for each author, each stage contains 3 novels with a maximum of 3 years between each publication. We propose several stylometric features to represent the novels in a vector space model. We use supervised learning algorithms to determine if by means of this stylometric-based representation is possible to identify to which stage of writing each novel belongs.

Keywords. Stylometry, writing style, authorship analysis.

1 Introduction

Currently, there is an exponential growth of digital information produced every day in the form of texts written in natural language, such as magazines, books, websites, newspapers, reports, etc. Nowadays, it is possible to process huge amounts of data from sources such as audio, video, images and text with machine learning algorithms. The authorship analysis studies [5], is one of the multiple application that can take advantage of this situation in order to turn the vast amount of data into practical and useful knowledge.

In authorship analysis, typical features used for text representation in the Vector Space Model (VSM) are words, Bag of Words (BoW) model [11], word n-grams [16, 22], character n-grams [7, 22], and syntactic n-grams [19]. The values of these features can be Boolean [15], tf-idf (term frequency-inverse document frequency), weights [12], or values based on probabilistic models [3]. Another popular statistical-based text representation are the stylometric features [23, 9], such as length of sentences, complexity of sentences, frequent words, spelling errors, etc.

In this paper, we aim to identify changes in the writing style of 7 authors of novels written in English using only stylometric features. Other text representations were evaluated for this corpus such as bag-of-word and n-grams (words, characters, POS tags, syntactic), in previous work [21]. From the machine learning perspective, this task can be viewed as a multi-class, single-label classification problem,

ISSN 1405-5546
doi: 10.13053/CyS-22-1-2882

when the automatic methods have to assign class labels (stages of writing for each author), to objects (text samples). This process requires a training and test set to generate a function that maps the input data with an output label. We examine various stylometric-based features, including lexical usage, punctuation and phraseology analysis; the evaluated machine-learning algorithms are liblinear and libSVM implementations of Support Vector Machines (SVM) and Logistic Regression (LG).

The identification of writing style changes have many applications, for example, it can be used to study early detection of diseases related memory loss and other cognitive abilities [6]. It is also important for the authorship attribution task when it is assumed that the writing style of the author is unique and stable, and can be detected in all his or her writing [1].

The paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we provide some characteristics of the corpus we used for the experiments. In Section 4, we describe the methodological framework. In Section 5, we present the obtained results. Finally, in Section 6, we draw some conclusions and point to possible directions of future work.

2 Related Work

In [8], the authors analyzed 14 Agatha Christie novels creating blocks of 10,000 words but they used the first 5 blocks of each novel. They used the wealth of vocabulary, n-grams of words and undefined words. They found that the wealth of vocabulary decreased as the author's age increased. The repeated phrases and the use of undefined words showed the opposite behavior.

The work [17], analyzed novels of different literary genres using LIWC characteristics. The results were correlated with the author's age at the time of writing the work. They concluded that the way people use language changes throughout their lives and that people show consistent changes in their language styles based on their age.

A study to measure the fidelity of an author in his writing style was carried out in [18], specifically the inter-tutorial variation at lexical, syntactic and semantic levels. A corpus of 20 opinion articles from 6 authors were compiled and the characteristics used were frequent words, slogans, the wealth of vocabulary, POS tags, content words, function words, sentence length, paragraph length, first level syntactic structures (chunks), word length, hapax legomena, distribution of signs of punctuation, adverbs (which end in mind), among others. The study was carried out using the analysis of variance (ANOVA). When evaluating the similarity of the texts through the 5 and 10 most frequent words, the results showed that the authors tend to use this type of words recurrently. The work concluded that the authors tend to maintain the same patterns of language instead of changing them with other options.

An analysis about the change of writing style of the Turkish authors Cetin Altan and Yasar Kemar was performed in [2]. A corpus compiled of 2 novels written in 1971 and 1998 by Kemar and 201 documents written between 1945 and 1991 by Yasar. The characteristics used were word length, length of word types and most frequent words. The principal component analysis (PCA), technique together with discriminant analysis and logistic regression were used to distinguish between old and new jobs. The results showed that the frequency and length of words in the most recent works was significantly higher than in the first works for both authors. To identify the membership of a block either to the old or the new period, the texts were divided into groups of 16 blocks. The works of Altan were identified with higher precision (suppose that this is due to the greater gap of time existing between the works of this author with respect to Kemal). In their conclusions, they indicate that when using discriminant analysis the success rate was of 98.96% and 84.38% for Altan and Kemal respectively.

In [6], were conducted experiments in order to identify writing style changes over the time in authors with Alzheimer's disease.They compiled a corpus gathering novels from 3 authors: Agatha Christie (15 novels), and Iris Murdoch (20 novels), affected by the Alzheimer disease and P.D James (15 novels), who was unaffected by the disease. The experiments were conducted using a set of well-known features in authorship attribution and

Table 1. Corpus description

Author	Initial Stage		Middle Stage		Final Stage	
	Year	Novel	Year	Novel	Year	Novel
Booth Tarkington (BT)	1899	Gentleman	1914	Penrod	1919	Ramsey
	1902	Vanrevels	1915	Turmoil	1921	Alice Adams
	1905	Canaan	1916	Seventeen	1922	Gentle Julia
Charles Dickens (CD)	1838	Nicholas Nickleby	1848	Dombey and Son	1859	Two Cities
	1838	Oliver Twist	1850	Copperdfield	1861	Expectations
	1841	Barnaby	1853	Bleak house	1865	Our mutual friend
Frederick Marryat (FM)	1830	The King's Own	1839	The panthom ship	1845	The Mission
	1831	Jacob Faithful	1839	A diary in America	1847	New Forrest
	1831	Newton Forster	1840	Olla Podrida	1848	The Little Savage
George MacDonald (GM)	1863	David Elginbrod	1873	Gutta Percha	1888	Electrical Lady.txt
	1864	Adela	1875	A double story	1891	Flight of Shadow
	1865	Alec Forbes	1876	Thomas Wingfold	1892	Hope of góspel
George Vaizey (GV)	1901	School Story	1908	Flaming June	1914	Cassandra
	1902	Pixie	1908	Big Game	1914	College Girl
	1902	Houseful of Girls	1910	Marriage	1915	Claire
Louis Tracy (LT)	1903	Wings of morning	1907	The captain	1912	Romance of NY
	1904	The revelers	1909	Inmortals	1916	The day of wrath
	1905	Disapperance	1909	The stoneway girl	1919	Mortimer fenley
Mark Twain (MT)	1869	Innocents Abroad	1883	Mississippi	1897	The Equator
	1872	Roughing It	1884	Huckleberry Finn	1905	What is man?
	1876	Tom Sawyer	1889	King Arthur	1906	Dollar

authorship verification, including frequent words, function words, characters, character n-grams, POS tags big-maps, POS tags entropy, most frequent words, among others. Both strategies unmasking technique and an SVM classifier were used to detect changes in authors' style. Experiments concluded that the proposed method was unable to detect the changes in the writing style caused by the disease and could not find any set of characteristics that would reliably discriminate the age of the authors.

3 Corpus

The corpus used in our study includes texts downloaded from the Project Gutenberg [20]. We selected books of native English speaking authors that had their literary production in a similar period. In this paper, all experiments were conducted for the corpus of sixty-three documents by seven authors. Table 1, shows the final corpus description.

In order to label the corpus with the writing time period, we first performed a chronological ranking of the novels by each of the seven authors based on the date of publication. Then we defined three writing stages (initial, middle and final), each stage contains three novels and there are at least two years of separation between the novels in each stage.

4 Methodology

In order to evaluate the performance of the classification models, we conformed testing sets with 3 novels (1 per stage), and training sets with 6 novels (the remaining 2 stage). There are in total nine novels for each author, thus twenty-seven different pairs of training-testing sets were obtained for each author. With this

Table 2. Testing sets for the author Booth Tarkington

1	Gentleman-Penrod-AliceAdms	10	Vanrevels-Penrod-AliceAdms	19	Canaan-Penrod-AliceAdams
2	Gentleman-Penrod-Julia	11	Vanrevels-Penrod-Julia	20	Canaan-Penrod-Julia
3	Gentleman-Penrod-Ramsey	12	Vanrevels-Penrod-Ramsey	21	Canaan-Penrod-Ramsey
4	Gentleman-Turmoil-AliceAdams	13	Vanrevels-Turmoil-AliceAdams	22	Canaan-Turmoil-AliceAdams
5	Gentleman-Turmoil-Julia	14	Vanrevels-Turmoil-Julia	23	Canaan-Turmoil-Julia
6	Gentleman-Turmoil-Ramsey	15	Vanrevels-Turmoil-Ramsey	24	Canaan-Turmoil-Ramsey
7	Gentleman-Seventeen-AliceAdams	16	Vanrevels-Seventeen-AliceAdams	25	Canaan-Seventeen-AliceAdams
8	Gentleman-Seventeen -Julia	17	Vanrevels-Seventeen -Julia	26	Canaan-Seventeen-Julia
9	Gentleman-Seventeen –Ramsey	18	Vanrevels-Seventeen- Ramsey	27	Canaan-Seventeen -Ramsey

experimental configuration, we ensure that all the novels are part of both sets (training and testing). Table 2 shows the twenty-seven testing sets for the author Booth Tarkington (BT). The training sets are not shown for space reasons.

We examined the performance of different stylometric features and machine learning algorithms. Stylometry is the analysis of style features that can be statistically quantified, such as sentence length, vocabulary diversity, and frequencies (of words, word forms, etc.). The stylometric has many practical applications, being one of the most popular the authorship attribution research, where the stylometric features are used as stylistic fingerprints for finding the author of anonymous or disputed documents.

We used a freely available Python library for obtaining the stylometric features [1] for all text samples. With the obtained features we built vector space models divided into three categories of stylometric analysis: phraseology analysis, punctuation analysis, and lexical usage analysis. The performance of each of the three feature sets was evaluated separately and in combinations. Table 3 shows the stylometric features that belong to each category of analysis. In the case of the lexical usage analysis, we used the list of stop words contained in the NLTK[2] library for Python.

We used the scikit-learn[3] implementation of the following machine learning classifiers: Logistic Regression and SVM (Liblinear and Libsvm implementations). These classification algorithms have proved to be among the best for text classification tasks [10, 13, 14].

Table 3. Features included in the three types of stylometric analysis

Analysis Type	**Features**
Phraseology (Phras.)	lexical diversity, mean word length, mean sentence length, stdev sentence length, mean paragraph length, stdev paragraph length, document length
Punctuation (Punct)	commas, semicolons, quotations, exclamations, colons, hyphens, double hyphens
Lexical Usage (Lex)	stop word list (a, the, of, in, etc.)

5 Results

We present the results in terms of accuracy achieved by the three stylometric analysis categories. The probability of assigning the correct class (writing stage), to a document at random is 33 %, this value is considered the baseline. First, we evaluated the classification algorithm that is better suited for this type of features. Figure 1, shows that in average the Logistic Regression algorithm obtained better results than the SVM-based algorithms, reaching a 5%, of difference in accuracy. However, the LibSVM algorithm achieved almost 10%, of improvement when using only the punctuation analysis features. For the individual results of each feature set, we only present here the results obtained with the Logistic Regression classifier.

[1] `https://github.com/jpotts18/stylometry`

[2] `http://www.nltk.org/`

[3] `http://scikit-learn.org/`

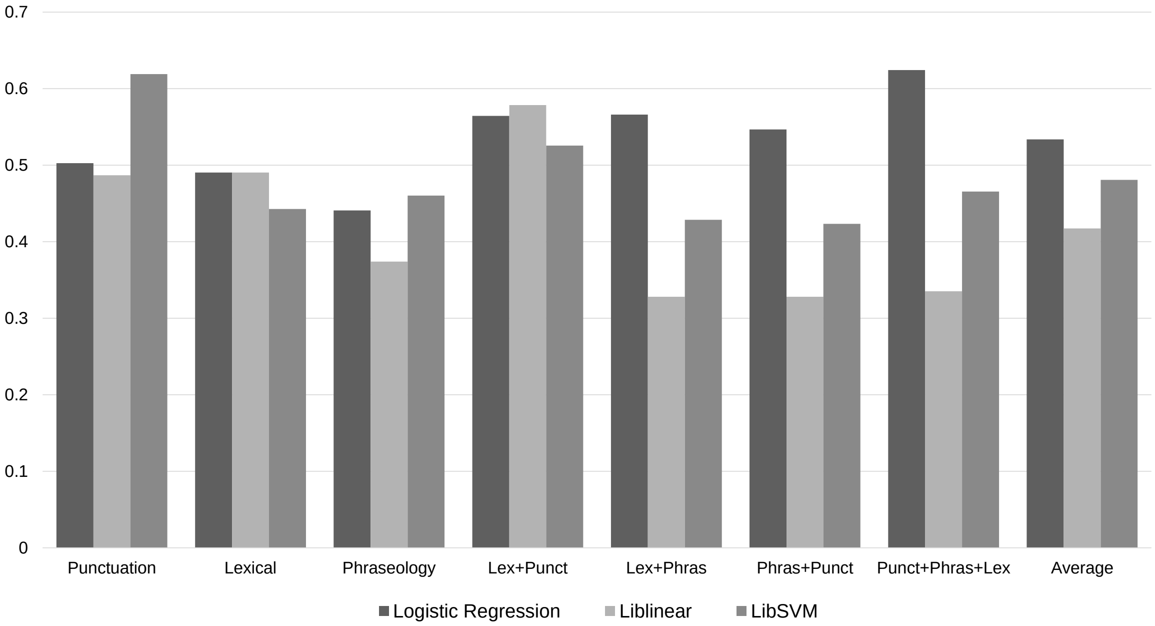

Fig. 1. Performace of the classification algorithms with each type of stylometric features

We evaluated the performance of the stylometric analysis individually and in combination for each author separately. Remember that we evaluate the classification performance with twenty-seven (27), testing sets and the final results represent the average obtained with the testing sets.

In Table 4, it can be observed that the highest classification accuracy is obtained in average when using the combination of all stylometric features. However, when the evaluation is performed individually, i.e. using only one type stylometric feature, the features within the punctuation analysis category yielded the best performance.

When analyzing the classification performance of the stylometric features on each author we observed that the classification model built on the punctuation-based features achieved very high performance. These models correctly classified the writing stage of a work above 70% of the times for two authors: Booth Tarkington (BT) and Mark Twain (MT). In the case of Charles Dickens (CD) and George MacDonald (GM), the combination of phraseology-and punctuation-based features obtained the best performance. The combination of all types of features obtained the best performance for Frederick Marryat (FM), George Vaizey (GV) and Louis Tracy (LT), classifying correctly the writing stage of the works 80% of the times on average.

6 Conclusion and Future Work

In this paper, we examined the performance of stylometric-based features for detecting writing style changes of seven authors of English novels. We defined three stages of writing for each author, with three novels each according to the publication date.

The obtained results indicate that the writing stage of a literary work can be identified with high accuracy (more than 70%), for four out of the seven evaluated authors using different the stylometric-based features. Furthermore, the classification models only failed in one of the authors (GM), where they obtained at most 56.8% of accuracy.

Table 4. Results in terms of accuracy with the Logistic Regression classifier for each author and type of stylometric analysis

Analysis Type	BT	CD	FM	GM	GV	LT	MT	Average
Punctuation	**0.827**	0.346	0.531	0.531	0.235	0.346	**0.704**	0.503
Lexical	0.667	0.407	0.457	0.272	0.481	0.630	0.519	0.490
Phraseology	0.185	0.481	0.543	0.333	0.580	0.667	0.296	0.441
Lex+Punct	0.741	0.481	0.580	0.556	0.556	0.519	0.519	0.564
Lex+Phras	0.457	0.593	0.753	0.469	0.605	0.753	0.333	0.566
Phras+Punct	0.235	**0.617**	0.654	**0.568**	0.593	0.852	0.309	0.547
Punct+Phras+Lex	0.519	0.605	**0.877**	0.543	**0.642**	**0.889**	0.296	**0.624**

We also found that for some authors, such as Louis Tracy, the writing style among the different stages can be identified with very high accuracy (88.9%). Even though the years of separation between each stage is at most three years. This seems to contradict some conclusions of the state-of-the-art where the authors found that the greater the gap of time existing between the novels the change is more evident [2].

This work serves as a baseline for more complex methods. One of the directions for future work will be to examine if is possible to identify the changes in writing style with other types of features such as documents embeddings [11, 4].

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT project 240844, CONACYT project FC-2016-01-2225, SNI, COFAA-IPN, SIP-IPN 20171813, 20171344, 20172008)

References

1. **Bagavandas, M. & Manimannan, G. (2008).** Style consistency and authorship attribution: A statistical investigation. *Journal of Quantitative Linguistics*, Vol. 15, No. 1, pp. 100–110.
2. **Can, F. & Patton, J. M. (2004).** Change of writing style with time. *Computers and the Humanities*, Vol. 38, No. 1, pp. 61–82.
3. **Croft, W. B., Turtle, H. R., & Lewis, D. D. (1991).** The use of phrases and structured queries in information retrieval. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pp. 32–45.
4. **Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., & Pinto, D. (2018).** Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, pp. 1–16.
5. **Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016).** Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors*, Vol. 16, No. 9, pp. 1374.
6. **Hirst, G. & Wei Feng, V. (2012).** Changes in style in authors with alzheimer's disease. *English Studies*, Vol. 93, No. 3, pp. 357–370.
7. **Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003).** N-gram-based author profiles for authorship attribution. *Proceedings of the conference pacific association for computational linguistics*, volume 3 of *PACLING'03*, pp. 255–264.
8. **Lancashire, I. & Hirst, G. (2009).** Vocabulary changes in agatha christie's mysteries as an indication of dementia: a case study. *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, pp. 8–10.
9. **López-Escobedo, F., Solorzano-Soto, J., & Sierra Martínez, G. (2016).** Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics*, Vol. 23, No. 2, pp. 154–176.
10. **Markov, I., Baptista, J., & Pichardo-Lagunas, O. (2017).** Authorship attribution in portuguese using

character n-grams. *Acta Polytechnica Hungarica*, Vol. 14, No. 3, pp. 59–78.

11. **Markov, I., Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., & Gelbukh, A. (2016).** Author profiling with doc2vec neural network-based document embeddings. *Mexican International Conference on Artificial Intelligence*, MICAI'16, Springer, pp. 117–131.

12. **Markov, I., Gómez-Adorno, H., & Sidorov, G. (2017).** Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. *Working Notes Papers of the CLEF*.

13. **Markov, I., Gómez-Adorno, H., Sidorov, G., & Gelbukh, A. (2017).** The winning approach to cross-genre gender identification in russian at rusprofiling 2017. *FIRE 2017 Working Notes*, FIRE'17, pp. 20–24.

14. **Markov, I., Stamatatos, E., & Sidorov, G. (2017).** Improving cross-topic authorship attribution: The role of pre-processing. *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'17.

15. **Mauldin, M. L. (1991).** Retrieval performance in ferret a conceptual information retrieval system. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'91, ACM, pp. 347–355.

16. **Mladenic, D. & Grobelnik, M. (1998).** Word sequences as features in text-learning. *Proceedings of the 17th Electrotechnical and Computer Science Conference*, ERK'98, pp. 145–148.

17. **Pennebaker, J. W. & Stone, L. D. (2003).** Words of wisdom: Language use over the life span. *Journal of personality and social psychology*, Vol. 85, No. 2, pp. 291.

18. **Pol, M. (2005).** A stylometry-based method to measure intra and inter-authorial faithfulness for forensic applications. *Workshop on Stylistic Analysis of Text for Information Access, ACM Press, Salvador, Bahia, Brazil*, SIGIR'05.

19. **Posadas-Durán, J.-P., Sidorov, G., Gómez-Adorno, H., Batyrshin, I., Mirasol-Mélendez, E., Posadas-Durán, G., & Chanona-Hernández, L. (2017).** Algorithm for extraction of subtrees of a sentence dependency parse tree. *Acta Polytechnica Hungarica*, Vol. 14, No. 3, pp. 79–98.

20. **Project Gutenberg (2018).** http://www.gutenberg.org. Accessed: January 1, 2018.

21. **Ríos-Toledo, G., Sidorov, G., Castro-Sánchez, N. A., & Posadas-Durán, J. P. (to appear).** Identification of changes in literary writing style using machine learning. *(submitted)*.

22. **Sanchez-Perez, M. A., Markov, I., Gómez-Adorno, H., & Sidorov, G. (2017).** Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. *International Conference of the Cross-Language Evaluation Forum for European Languages*, CLEF'17, Springer, pp. 145–151.

23. **Zheng, R., Li, J., Chen, H., & Huang, Z. (2006).** A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the Association for Information Science and Technology*, Vol. 57, No. 3, pp. 378–393.

Article received on 07/10/2017; accepted on 15/01/2018.
Corresponding author is Grigori Sidorov.