Author Identification using Writer-Dependent and Writer-Independent Strategies

Daniel Pavelec[†], Edson Justino[†], Leonardo V. Batista[‡], and Luiz S. Oliveira[†]

[†]Pontifícia Universidade Católica do Paraná (PUCPR)
Programa de Pós-Graduacão em Informática
{pavelec,justino,soares}@ppgia.pucpr.br

[‡]Federal University of Paraíba (UFPB)
Programa de Pós-Graduacão em Informática
leonardo@di.ufpb.br

ABSTRACT

In this work we discuss author identification for documents written in Portuguese. Two different approaches were compared. The first is the writer-independent model which reduces the pattern recognition problem to a single model and two classes, hence, makes it possible to build robust system even when few genuine samples per writer are available. The second is the personal model, which very often performs better but needs a bigger number of samples per writer. We also introduce a stylometric feature set based on the conjunctions and adverbs of the Portuguese language. Experiments on a database composed of short articles from 30 different authors and Support Vector Machine (SVM) as classifier demonstrate that the proposed strategy can produced results comparable to the literature.

Categories and Subject Descriptors

H.4 [**Pattern Recognition**]: Miscellaneous; D.2.8 [**Doc. Engineering**]: Stylometry—*document analysis*

Keywords

Author Identification, Stylometry

1. INTRODUCTION

There exists a long history of linguistic and stylistic investigation into author identification which goes back to the late nineteenth century, with the pioneering studies of Mendenhall [11] and Mascol [10] on distributions of sentence and word lengths in works of literature and the gospels of the New Testament. Modern work in author identification was preceded by Mosteller and Wallace in the 1960s, in their seminal study The Federalist Papers [13]. All these have been motivated by the fact that we usually leave indicative of authorship in our writings due to the fact that we have distinctive ways of writing [12].

In recent years, practical applications for author identification have grown in several different areas such as, criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (mining email content). Chaski [5] points out that in the investigation of certain crimes involving digital evidence, when a specific machine is identified as the source of documents, a legitimate issue is to identify the author that produced the documents, in other words, "Who was at the keyboard when the relevant documents were produced?".

In order to identify the author, one must extract the most appropriate features to represent the style of an author. In this context, the stylometry (application of the study of linguistic style) offers a strong support to define a discriminative feature set. The literature shows that several stylometric features that have been applied include various measures of vocabulary richness and lexical repetition based on word frequency distributions. As observed by Madigan et al [9], most of these measures, however, are strongly dependent on the length of the text being studied, hence, are difficult to apply reliably. Many other types of features have been tried out, including word class frequencies [7, 1], syntactic analysis [3], word collocations [16], grammatical errors [8], word, sentence, clause, and paragraph lengths [2].

To deal with the problem of author identification usually a writer-specific model (also known as personal model) is considered. It is based on two different classes, ω_1 and ω_2, where ω_1 represents authorship while ω_2 represents forgery. The main drawbacks of the writer-specific approach are the need of learning the model each time a new author should be included in the system and the great number of genuine samples of texts necessary to build a reliable model. An alternative to this strategy is the writer-independent approach. It uses the dissimilarity representation [14] and can be defined as writer-independent approach as the number of models does not depend on the number of writers. In this context, it is a global model by nature, which reduces the pattern recognition problem to a global model with two classes, consequently, makes it possible to build robust author identification systems even when few genuine samples per author are available.

SAC'08 March 16-20, 2008, Fortaleza, Ceará, Brazil

In this work we discuss the two aforementioned approaches for writer identification. We also propose a stylometric feature set for the Portuguese language, which is based on conjunctions and adverbs. Comprehensive experiments on a database composed of short articles and Support Vector Machine (SVM) as classifier demonstrate the advantages and drawbacks of each strategy and also that both can produce results comparable to the literature.

The remaining of this paper is divided as follows: Section 2 introduces the basic concepts of forensic stylistics and describes the linguistic features used in this work. Section 2.2 describes the basic concepts of the SVM. Section 3 describes how both writer-independent and writer-dependent approaches work. Section 3.1 presents the database used in this work. Section 4 describes both writer-dependent and writer-independent methods for author identification while Section 5 reports the experimental results. Finally, Section 6 concludes this work.

2. FORENSIC STYLISTICS

Forensic stylistics is a sub-field of forensic linguistics and it aims at applying stylistics to the context of author identification. The stylistic is based on two premisses:

- Two writers (same mother-tongue) do not write in the same way.
- The writer does not write in the same way all the time.

The stylistic can be classified into two different approaches: qualitative and quantitative.

The qualitative approach assesses errors and personal behavior of the authors, also known as idiosyncrasies, based on the examiner's experience. According to Chaski [5], this approach could be quantified through databasing, but until now the databases which would be required have not been fully developed. Without such databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias. In this vein, Koppel and Schler [8] proposed the use of 99 error features to feed different classifiers such as SVM and decision trees. The best result reported was about 72% of recognition rate.

The second approach, which is very often refereed as stylometry, is quantitative and computational, focusing on readily computable and countable language features, e.g. word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths. It uses standard syntactic analysis from the dominant paradigm in theoretical linguistics over the past forty years. Examples of this approach can be found in Tambouratzis et al [17] and Chaski [5]. Experimental results show that usually this approach provides better results than the qualitative one. For this reason we have chosen this paradigm to support our work.

2.1 Linguistic Features

The literature suggests many linguistic features to be used for author identification. In [4], Chaski discusses about the differences between scientific and replicable methods for author identification. Scientific methods are based on empirical, testable hypotheses, and the use of these methods can be done by anyone, i.e., it is not dependent on a special talent. In the same work, nine empirical hypotheses that have been used to identify authors in the past are reported: Vocabulary Richness, Hapax Legomena, Readability Measures, Content Analysis, Spelling Errors, Grammatical Errors, Syntactically Classified Punctuation, Sentential Complexity, Abstract Syntactic Structures.

Table 1: Conjunctions of the Portuguese language

Group	Conjunctions (in Portuguese)
Coordinating additive	e, nem, mas também, senão também, bem como, como também, mas ainda.
Coordinating adversative	porém, todavia, mas, ao passo que, não obstante, entretanto, senão, apesar disso, em todo caso contudo, no entanto
Coordinating conclusive	logo, portanto, por isso, por conseguinte.
Coordinating explicative	porquanto, que, porque.
Subordinating comparative	tal qual, tais quais, assim como, tal e qual, tão como, tais como, mais do que, tanto como, menos do que, menos que, que nem, tanto quanto, o mesmo que, tal como, mais que.
Subordinating conformative	consoante, segundo, conforme.
Subordinating concessive	embora, ainda que, ainda quando, posto que, por muito que, se bem que, por menos que, nem que, dado que mesmo que, por mais que.
Subordinating conditional	se, caso, contanto que, salvo que, a não ser que, a menos que
Subordinating consecutive	de sorte que, de forma que, de maneira que, de modo que, sem que
Subordinating final	para que, fim de que
Subordinating proportional	a proporção que, quanto menos, quanto mais a medida que.

Vocabulary Richness is given by the ratio of the number of distinct words (type) to the number of total words (token). Hapax Legomena is the ratio of the numbers of words occurring once (Hapax Legomena) to the total number of words. Readability Measures compute the supposed complexity of a document, and are calculations based on sentence length and word length. Content Analysis classifies each word in the document by semantic category, and statistically analyze the distance between documents. Spelling Errors quantifies the misspelled words. Prescriptive Grammatical Errors test errors such as sentence fragment, run-on sentence, subject-verb mismatch, tense shift, wrong verb form, and missing verb. Syntactically Classified Punctuation takes into account end-of-sentence period, comma sep-

arating main and dependent clauses, comma in list, etc. Finally, Abstract Syntactic Structures computationally analyzes syntactic patterns. It uses verb phrase structure as a differentiating feature.

In this work we propose the use of conjunctions and adverbs of the Portuguese language. Just like other language, Portuguese has a large set of conjunctions that can be used to link words, phrases, and clauses. Table 1 describes all the Portuguese conjunctions we have used in this work.

Such conjunctions can be used in different ways without modifying the meaning of the text. For example, the sentence "Ele *tal qual* seu pai" (He is like his father), could be written is several different ways using other conjunctions, for example, "Ele *tal e qual* seu pai", "Ele *tal como* seu pai", "Ele *que nem* seu pai", "Ele *assim como* seu pai". The way of using conjunctions is a characteristic of each author, and for this reason we decided to use them in this work.

To complete the feature set, we have used adverbs of the Portuguese language. An adverb can modify a verb, an adjective, another adverb, a phrase, or a clause. Authors can use it to indicate manner, time, place, cause, or degree and answers questions such as "how, "when, "where, "how much. Table 2 reports the list of 94 adverbs we have used in this work.

Table 2: Adverbs of the Portuguese language

Group	Conjunctions (in Portuguese)
Place	aqui, ali, aí, cá, lá, acolá, além, longe, perto, dentro, adiante, defronte, onde, acima, abaixo, atrás, em cima, de cima, ao lado, de fora, por fora.
Time	hoje, ontem, amanhã, atualmente, sempre, nunca, jamais, cedo, tarde, antes, depois, j, agora, então, de repente, hoje em dia.
Affirmation	certamente, com certeza, de certo, realmente, seguramente, sem dúvida, sim
Intensity	ainda, apenas, de pouco, demais, mais, menos, muito, pouca, pouco, quase, tanta, tanto
Negative	absolutamente, de jeito nenhum, de modo algum, não, tampouco
Subordinating concessive	embora, ainda que, ainda quando, posto que, por muito que, se bem que, por menos que, nem que, dado que mesmo que, por mais que.
Quantity	todo, toda
Mode	assim, depressa,bem, devagar, face a face, facilmente, frente a frente, lentamente, mal, rapidamente, algo, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo

2.2 Author Identification with SVM

As stated before two different models for author identification are the subject of this work. In both strategies binary classifiers fit quite well. For the global approach just one model should be built while for the personal approach one binary model for each author is necessary. In light of this, Support Vector Machine (SVM) [18] seems quite suitable since it was originally developed to deal with problems with two classes. Moreover, SVM is tolerant to outliers and perform well in high dimensional data.

One of the limitations with SVMs is that they do not work in a probabilistic framework. There is several situations where would be very useful to have a classifier producing a posterior probability $P(class|input)$. In our case, particulary, we are interested in estimation of probabilities because we want to try different fusion strategies like Max, Min, Average, and Median.

Due to the benefits of having classifiers estimating probabilities, many researchers have been working on the problem of estimating probabilities with SVM classifiers. The one suggested by Platt [15] uses a slightly modified logistic function, defined as:

$$P(y = 1|f(x)) = \frac{1}{1 + \exp(Af(x) + B))} \quad (1)$$

It has two parameters trained discriminatively, rather one parameter estimated from a tied variance. The parameters A and B of Equation 1 are found by minimizing the negative log likelihood of the training data, which is a cross-entropy error function.

3. WRITER-DEPENDENT VS WRITER-INDEPENDENT

The writer-dependent or personal model is based on one model per author. Usually it yields good results but its drawback lies in the fact that for each new author a new model should be built. Another important issue in this strategy is that usually a considerable amount of data is necessary to train a reliable model. It can be implemented using either one-against-all or pairwise strategy. This kind of approach has been largely used for signature verification.

An alternative to the personal approach is the global approach or writer-independent model. It is based on the forensic questioned document examination approach and classifies the writing, in terms of authenticity, into genuine and forgery, using for that one global model. In the case of author identification, the experts use a set of n genuine articles $Sk_i, (i = 1, 2, 3, \ldots, n)$ as references and then compare each Sk with a questioned sample Sq. The idea is to verify the discrepancies among Sk and Sq. Let V_i be the stylometric feature vectors extracted from the reference articles and Q the stylometric feature vector extracted from the questioned article. Then, the dissimilarity feature vectors $Z_i = \|V_i - Q\|_2$ are computed to feed n different instances of the classifier C, which provide a partial decision. The final decision D depends on the fusion of these partial decisions, which are usually obtained through the majority vote rule. Figure 1 depicts the global approach.

Note that when a dissimilarity measure is used, the components of the feature vector Z tends to be close to 0 when both the reference Sk and the questioned Q comes from the same author. Otherwise, the feature vector Z tend to be far

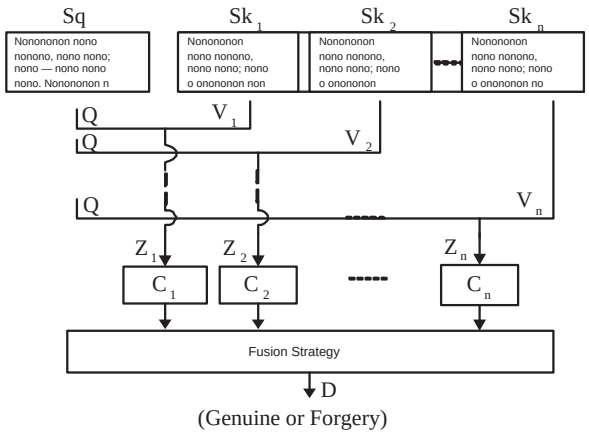

Figure 1: Architecture of the global approach.

from 0.

3.1 Database

To build the database we have collected articles available in the Internet from 30 different people with profiles ranging from sports to economics. Our sources were two different Brazilian newspapers, *Gazeta do Povo* (http://www.gazetadopovo.com.br) and *Tribuna do Paran* (http://www.paranaonline.com.br). We have chosen 30 short articles from each author. The articles usually deal with polemic subjects and express the author's personal opinion. In average, the articles have 600 tokens and 350 Hapax.

One aspect worth of remark is that this kind of articles can go through some revision process, which can remove some personal characteristics of the texts. Figure 2 depicts an example of the article of our database.

COLUNISTAS

ANTÔNIO ERMÍRIO DE MORAES

Um programa que jamais deve ser esquecido

São Paulo é uma bela cidade que, infelizmente, está sendo atacada por pichadores de paredes que praticam um verdadeiro vandalismo contra o patrimônio público e privado. A eles se juntam os que utilizam as praças públicas para dormir, comer, lavar roupa e fazer necessidades. Mais triste é ver a proliferação de menores e adolescentes nos cruzamentos das ruas e pedir esmolas e vender quinquilharias.

O caso dos menores é muito triste por dois motivos. Primeiro, há a dimensão humana. O lugar de menores é na escola, em casa e em ambientes de lazer sadio. É aí que eles aprendem os valores fundamentais do convívio civilizado. Deixá-los nas ruas é submetê-los a um abandono que certamente irá deformar seu caráter.

Figure 2: An example of an article used in this work.

4. IMPLEMENTATION

This section describes how both strategies have been implemented. In both cases we have used a feature vector of 171 components, which is composed of 77 conjunctions and 94 adverbs. In order to extract the features, first the text is segmented into tokens. Spaces and end-of-line characters are not considered. All hyphenized words are considered as two words. In the example, the sentence "*eu vou dar-te um pula-pula e também dar-te-ei um beijo, meu amor*!" has 16 tokens and 12 Hapax. Punctuation, special characters, and numbers are not considered as tokens. There is no distinction between upper case and lower case.

4.1 Writer-Dependent

There are two basic approaches to solve q-class problems with SVMs: pairwise and one-against-others. In this work we have used the former, which arranges the classifiers in trees, where each tree node represents a SVM. For a given test sample, it is compared with each two pairs, and the winner will be tested in an upper level until the top of the tree. In this strategy, the number of classifiers we have to train is $q(q-1)/2$.

From the database described previously, we have used 20 authors ($q = 20$, consequently 190 models). From each author 10 documents were used for training and 15 documents for testing.

4.2 Writer-Independent

Differently of the writer-dependent approach, this strategy consists in training just one global model which should discriminate between author (ω_1) and not author (ω_2). To generate the samples of ω_1, we have used three articles (A_i) for each author. Based on the concept of dissimilarity, we extract features for each article and then compute the dissimilarities among them as shown in Section 3. In this way, for each author we have 10 feature vectors, summing up 100 samples for training (10 authors). The samples of ω_2 were created by computing the dissimilarities of the articles written by different authors, which were chosen randomly. As stated before, the proposed protocol takes into consideration a set of references (Sk). In this case we have used 20 authors (the same 20 used for the writer-dependent), five articles per author as references and 15 as questioned (Sq - testing set).

Following the protocol introduced previously, a feature vector composed of 171 components is extracted from the questioned (Sq) and references (Sk_i) documents as well. This produces the aforementioned stylometric feature vectors V_i e Q. Once those vectors are generated, the next step consists in computing the dissimilarity feature vector $Z_i = \|V_i - Q\|_2$, which will feed the SVM classifiers. Since we have five ($n = 5$) reference images, the questioned image Sq will be compared five times (the SVM classifier is called five times), yielding five votes or scores. When using discrete SVM, it produces discrete outputs $\{-1, +1\}$, which ca be interpreted as votes. To generate scores, we have used the probabilistic framework described in Section 2.2. Finally, the final decision can be taken based on different fusion strategies, but usually majority voting is used.

5. RESULTS

In this section we report the experiments we have performed. In both strategies, different parameters and kernels for the SVM were tried out but the better results were yielded using a linear kernel.

Considering the writer-dependent model, the best result we got was 83.2% of recognition rate. As mentioned previously, few works have been done in the field of author identification for documents written in Portuguese. For this reason is quite difficult to make any kind of direct comparison. To the best of our knowledge, the only work dealing with author identification for documents written in Portuguese was proposed by Coutinho et al [6]. In this work the authors

extract features using a compression algorithm and achieve a recognition rate of 78%. However, the size of the texts used for feature extraction is about 10 times bigger.

As one could observe, the main disadvantage of the writer-dependent model is the huge number of models necessary. This approach is unfeasible as the number of authors gets bigger. One alternative to surpass this problem is the writer-independent model, which does not depend on the number of author. Using this approach the best result we got was 75.1%. Contrary to the writer-dependent approach where we have used a feature vector composed of conjunctions and adverbs, here the best results were produced using only 77 conjunction features. Table 5 summarizes the results.

Table 3: Results on the test set composed of 200 documents from 20 different authors

Strategy	Rec. Rate (%)
Writer-dependent	83.2%
Writer-independent	75.1%

In spite of the fact that the writer-independent approach achieves worse results, we argue that it should be considered as an alternative because of its lower computational complexity. Besides, we believe that the writer-independent can be improved if we investigate different types of features.

6. CONCLUSION

In this paper we have compared two different strategies for author identification using a feature set based on conjunctions and adverbs of the Portuguese language. We could observe that the writer-dependent method achieves better results but at an elevated computation cost. On the other hand, the writer-independent is quite simple as strategy and has a very accessible cost, but it has a bigger error rate. If the application has few writers, the writer-dependent should the strategy to be considered. But if the number of writes gets bigger, writer-independent should be taken into account as alternative.

Comprehensive experiments on a database composed of short articles from 30 different authors demonstrate that both strategies produce results comparable to the literature. As future work, we plan to increase the database and define new features so that the overall performance of the system could be improved.

Acknowledgements

This research has been supported by The National Council for Scientific and Technological Development (CNPq) grant 475645/2004-9.

7. REFERENCES

[1] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3), 2003.

[2] S. Argamon, M. Saric, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In *ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[3] H. Baayen, H. van Halteren, and F. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131, 1666.

[4] C. Chaski. A daubert-inspired assessment of current techniques for language-based author identification. Technical Report 1098, ILE Technical Report, 1998.

[5] C. E. Chaski. Who is at the keyboard. authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1), 2005.

[6] B. C. Coutinho, L. M. Macedo, A. Rique-JR, and L. V. Batista. Atribuio de autoria usando PPM. In *XXV Congress of the SBC*, pages 2208–2217, 2004.

[7] R. S. Forsyth and D. I. Holmes. Feature finding for text classfication. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[8] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[9] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. Author identification on the large scale. In *Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA)*, 2005.

[10] C. Mascol. Curves of pauline and pseudo-pauline style i. *Unitarian Review*, 30:453–460, 1888.

[11] T. Mendenhall. The characteristic curves of composition. *Science*, 214:237–249, 1887.

[12] A. Morton. *Literary Detection*. Charles Scribners Sons, 1978.

[13] F. Mosteller and D. L. Wallace. Inference and disputed authorship: The federalist. In *Series in behavioral science: Quantitative methods edition*. Addison-Wesley, 1964.

[14] E. Pekalska and R. P. W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition*, 23:943–956, 2002.

[15] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola et al, editor, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[16] F. Smadja. Lexical co-occurrence: The missing link. *Journal of the Association for Literary and Linguistic Computing*, 4(3), 1989.

[17] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, and D. Tambouratzis. Discriminating the registers and styles in the modern greek language – part 2: Extending the feature vector to optimize author discrimination. *Literary and Linguistic Computing*, 19(2):221–242, 2004.

[18] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc, 1995.