Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 – August 5, 2011

Selecting Syntactic Attributes for Authorship Attribution

Paulo Varela, Edson Justino, and Luiz S. Oliveira

Abstract— **In this work we present a methodology to select syntactic attributes for authorship attribution. The approach takes into account a multi-objective genetic algorithm and a Support Vector Machine classifier and it operates in a wrapper mode. Through a series of comprehensive experiments on a database composed of 3000 short articles written in Portuguese we show that the proposed methodology is able to provide a concise subset of attributes, which increases the recognition rate in about 15 percentage points.**

I. Introduction

AUthorship attributtion can be defined as the task of inferring characteristics of a document's author from the textual characteristics of the document itself. Such an analysis can be performed either on a piece of handwritten text or on a digital document. In the first case the experts would look for idiosyncratic loop of an "e", slant of an "l", and other graphometric features that reliably characterize the writer. This problem is referred in the literature as writer recognition. In the second case, such features are not available since the document is in digital format. The challenge here is to estimate how similar two documents are from each other, based on patterns of linguistic behavior in documents of known and unknown authorship. This is known in the literature as authorship attribution or authorship analysis, which is the focus of this work.

In recent years, practical applications for authorship attribution 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 [1] 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 linguistic style offers a strong support to define a discriminative feature set. The linguistics, which can be defined as the scientific study of the human language, can be broadly divided into stylistics (variations of the language within a context), syntax (how language combines words to form grammatical sentences), lexicology (the set of words), morphology (internal structure of words), and grammar (structural rules that govern the composition of sentences). From the forensic perspective, each area described above may contribute to the formation of a minimum set of robust and mutually independent attributes, which are able to establish the authorship of a given text.

Stylistic attributes strongly depend on the theme of the text in question, which disfavors the distinction of different authors dealing with the same subject. Lexicographical attributes can aid in the establishment of the lexical richness of the author, however, disfavors the distinction of authors who possess a wealth of similar vocabulary.

The morphology can assist in identifying variations of the spelling of words written by the author, common in certain types of informal texts such as e-mails, however, in formal texts these variabilities are less or never observable.

The syntax and grammar have a complex set of attributes that together are capable of establishing standards for authorship attribution independent of the subject and the type of text, formal or informal. The richness and diversity of vocabulary are intrinsic features for these two classes of attributes. Two subclasses of attributes stand out in the syntactic, they are, variable and invariable.

The problem of using such attributes for authorship attribution lies in the huge number of available features. With this in mind, in this work we present a methodology to select the most discriminative syntactic attributes of the Portuguese language for the task of authorship attribution. The initial feature set is a bag of 408 words composed of variable (verbs and nouns) and invariable (conjunctions and adverbs) syntactic attributes. The proposed methodology is a wrapper approach that uses a multi-objective genetic algorithm [4] to generate a set of alternative solutions and a validation set to indicate the best accuracy/complexity tradeoff. The classification accuracy is supplied by a Support Vector Machine classifier.

Comprehensive experimental results on a database composed of 3000 documents from 100 different authors writing about 10 distinct subjects show that the number of attributes can be considerably reduced while improving the performance of the system in more than 15 percentage points.

II. Authorship Attribution with SVM

To deal with the problem of author attribution usually an author-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 author-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 text necessary to build a reliable model. An alternative to this strategy is the author-independent approach. It uses the dissimilarity representation [8] and can

Paulo Varela and Edson Justino are with PUCPR, Curitiba, PR, Brazil, and Luiz S. Oliveira is with the Department of Informatics of UFPR, Curitiba, PR, Brazil. (email: justino@ppgia.pucpr.br, lesoliveira@inf.ufpr.br

This work was supported by the CNPq, project grant #306358/2008-5.

978-1-4244-9637-2/11/$26.00 ©2011 IEEE

be defined as an author-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.

In light of this, Support Vector Machine (SVM) [14] 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 are several situations where would be very useful to have a classifier producing a posterior probability $P(class|input)$. In our case, particularly, 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 [9] uses a slightly modified logistic function, defined as:

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

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.

A. The Author-Independent Approach

The author-independent approach 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 attribution, 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 tends to be far from 0. Of course this is totally true under favorable conditions. As any other feature representation, the dissimilarity feature vector can be affected by the intra-writer variability. Such a variability could generate values far from zero when measuring the dissimilarity of genuine writers.

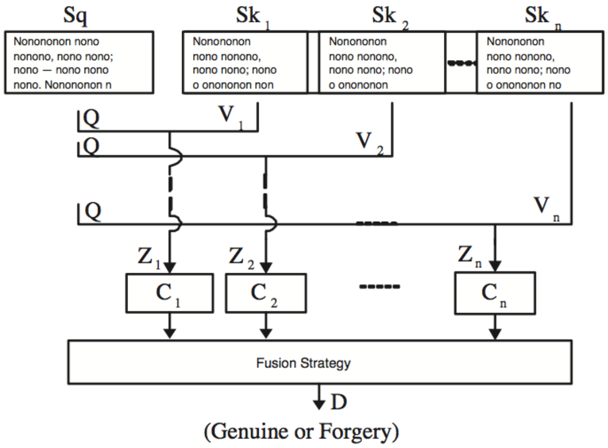


Fig. 1. Architecture of the global approach.

III. Multi-Objective Feature Selection

In the context of practical applications feature selection presents a multi-criterion optimization function, e.g. number of features and accuracy of classification. Genetic algorithms offer a particularly attractive approach to solve this kind of problems since they are generally quite effective in rapid global search of large, non-linear and poorly understood spaces.

A general multi-objective optimization problem consists of a number of objectives and is associated with a number of inequality and equality constraints. Solutions to a multi-objective optimization problem can be expressed mathematically in terms of nondominated points, i.e., a solution is dominant over another only if it has superior performance in all criteria. A solution is said to be Pareto-optimal if it cannot be dominated by any other solution available in the search space

A common difficulty with multi-objective optimization problem is the conflict between the objectives. In general, none of the feasible solutions allow simultaneous optimal solutions for all objectives. Thus, mathematically the most favorable Pareto-optimum is the solution that offers the least objective conflict. In order to find such solutions, classical methods scalarize the objective vector into one objective

The simplest of all classical techniques is the weighted sum method. It aggregates the objectives into a single and parameterized objective through a linear combination of the objectives. However, setting up an appropriate weight vector also depends on the scaling of each objective function. It is likely that different objectives take different orders of magnitude. When such objectives are weighted to form a composite objective function, it would be better to scale them appropriately so that each has more or less the same order or magnitude. Moreover, the solution obtained through this strategy largely depends on the underlying weight vector.

A. Pareto-based Approach

In order to overcome such difficulties, Pareto-based evolutionary optimization has become an alternative to classical techniques such as weighted sum method. This approach was first proposed by Goldberg in [2] and it explicitly uses Pareto dominance in order to determine the reproduction probability of each individual. Basically, it consists of assigning rank 1 to the nondominated individuals and removing them from contention, then finding a new set of nondominated individuals, ranked 2, and so forth.

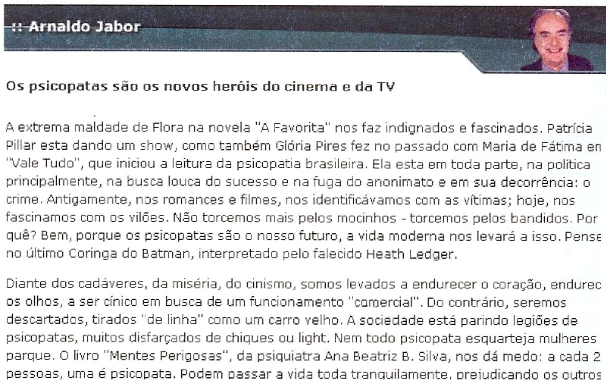
Pareto-based ranking correctly assigns all nondominated individuals the same fitness, however, this does not guarantee that the Pareto set be uniformly sampled. In order to avoid such a problem, Goldberg and Richardson in [1] pro- pose the additional use of fitness sharing. The main idea behind this is that individuals in a particular niche have to share the available resources. The more individuals are located in the neighborhood of a certain individual, the more its fitness value is degraded.

In this work, we have used the Nondominated Sorting Genetic Algorithm NSGA II proposed by Deb in [4]. The idea behind NSGA is that a ranking selection method is used to emphasize good points and a niche method is used to maintain stable subpopulations of good points. It varies from simple genetic algorithm only in the way the selection operator works. The crossover and mutation remain as usual. Before the selection is performed, the population is ranked on the basis of an individual's nondomination. The nondominated individuals present in the population are first identified from the current population. Then, all these individuals are assumed to constitute the first nondominated front in the population and assigned a large dummy fitness value. The same fitness value is assigned to give an equal reproductive potential to all these nondominated individuals.

In order to maintain the diversity in the population, these classified individuals are then shared with their dummy fitness values. Sharing is achieved by performing selection operation using degraded fitness values obtained by dividing the original fitness value of an individual by a quantity proportional to the number of individuals around it. Thereafter, the population is reproduced according to the dummy fitness values. Since individuals in the first front have the maximum fitness value, they get more copies than the rest of the population. The efficiency of NSGA lies in the way multiple objectives are reduced to a dummy fitness function using nondominated sorting procedures. More details about NSGA can be found in [4].

IV. Database

To build the database we have collected articles available in the Internet from 100 different authors which were uniformly distributed into 10 different subjects: Miscellaneous, Law, Economics, Sports, Gastronomy, Literature, Politics, Health, Technology, and Tourism. Our sources were 15 Brazilian newspapers located all over the country. Figure 2 shows an example of an article of the database.



:: Arnaldo Jabor

Os psicopatas são os novos heróis do cinema e da TV

A extrema maldade de Flora na novela "A Favorita" nos faz indignados e fascinados. Patrícia
Pillar esta dando um show, como também Glória Pires fez no passado com Maria de Fátima en
"Vale Tudo", que iniciou a leitura da psicopatia brasileira. Ela esta em toda parte, na política
principalmente, na busca louca do sucesso e na fuga do anonimato e em sua decorrência: o
crime. Antigamente, nos romances e filmes, nos identificávamos com as vítimas; hoje, nos
fascinamos com os vilões. Não torcemos mais pelos mocinhos - torcemos pelos bandidos. Por
quê? Bem, porque os psicopatas são o nosso futuro, a vida moderna nos levará a isso. Pense
no último Coringa do Batman, interpretado pelo falecido Heath Ledger.

Diante dos cadáveres, da miséria, do cinismo, somos levados a endurecer o coração, endurec
os olhos, a ser cínico em busca de um funcionamento "comercial". Do contrário, seremos
descartados, tirados "de linha" como um carro velho. A sociedade está parindo legiões de
psicopatas, muitos disfarçados de chiques ou light. Nem todo psicopata esquarteja mulheres
parque. O livro "Mentes Perigosas", da psiquiatra Ana Beatriz B. Silva, nos dá medo: a cada 2
pessoas, uma é psicopata. Podem passar a vida toda tranquilamente, prejudicando os outros

Fig. 2. An example of an artice of the database

We have chosen 30 short articles from each author, thus summing up 3000 pieces of documents. The articles usually deal with polemic subjects and express the authors personal opinion. In average, the articles have 600 tokens (words) and 350 Hapax (words occurring once). 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. Besides, authorship attribution using short articles poses an extra challenge since the number of features that can be extracted are directly related to the size of the text.

For our experiments, the database was divided into training (20%), validation (20%), searching (20%) and testing (40%) sets. Training and validation were used during the learning phase, the searching set was used to compute the fitness during the search, and the testing set was used for final testing.

V. Features

Two different approaches can be used for authorship attribution: 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 [1], 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.

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 [10], Chaski [1], and Pavelec et al [7]. 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.

As stated before, four different types of syntactic features were selected for this study. The first set of feature contains 77 conjunctions. Just like other language, Portuguese has a large set of conjunctions that can be used to link words, phrases, and clauses. Such conjunctions can be used in different ways without modifying the meaning of the text. For example, the sentence "Ele é *tal qual* seu pai" (He is like his father), could be written is several different ways using other conjunctions, for example, "Ele é *tal e qual* seu pai", "Ele é *tal como* seu pai", "Ele é *que nem* seu pai", "Ele é *assim como* seu pai". The way of using conjunctions is a characteristic of each author. Table I describes all the Portuguese conjunctions we have used in this work.

TABLE I
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, senão, entretanto não obstante, 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, nem que por muito que, e bem que, por menos 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

The second feature set contains 94 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 II reports the list of adverbs we have used in this work.

As discussed before, both conjunctions and adverbs are invariable syntactic features. To complement the feature set, we have added two classes of variable syntactic features, namely, verbs and pronouns. In the case of the verbs, we have used the 50 most used verbs of the Brazilian Portuguese language in three different forms: infinitive, gerund, and past participle. This sums up 150 attributes. Finally, 87 pronouns were added to our bag of words. Table III and IV show the verbs and pronouns, respectively.

TABLE II
ADVERBS OF THE PORTUGUESE LANGUAGE

Group	Adverbs (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, já, 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, algo, facilmente, frente a frente, lentamente, mal, rapidamente, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo

TABLE III
VERBS OF THE PORTUGUESE LANGUAGE

verbs (in Portuguese)
escrever, falar, jogar, andar, ver, ser, cantar, pular, ler, ter, achar colar, estar, dizer, dar, escolher, fechar, entender, fazer, trocar, abrir, acabar, declarar, completar, visitar, encerrar, comer, beber, pensar, possuir, atingir, melhorar, achar, realizar, haver, viver, aplicar, gerar, melhorar, pagar, distribuir, ligar, usar, projetar, desenvolver, poder, implantar, trazer, iniciar, efetuar

VI. EXPERIMENTS AND DISCUSSION

Before discussing feature selection, our first experiment consisted in training a baseline classifier using the 408 features available. Different kernels and parameters for the SVM classifier were tried out but in all experiments the linear kernel provided the best results. Therefore, the SVM with a linear kernel was used in all experiments.

After training the classifier with all the features, the best recognition rate on the testing set was 58%. We also have trained four different classifiers, one for each feature set. None of these classifiers were able to overcome the baseline classifier.

Regarding the feature selection, the NSGA was based on bit representation, one-point crossover, bit-flip mutation, roulette wheel selection, and elitism which is implemented using a generational procedure. The following parameter settings were employed: Population size = 128, Number of generations = 1000, Probability of crossover = 0.8, Probability of mutation = 1/number of features. In order to define the probabilities of crossover and mutation, we have used the one-max problem, which is probably the most frequently-used test function in research on genetic algorithms because of its simplicity.

TABLE IV

PRONOUNS OF THE PORTUGUESE LANGUAGE

Group	Pronouns (in Portuguese)
Relatives	quem, o qual,a qual, os quais, as quais,onde, em que, quanto, quanta, quantos, quantas, cujo, cuja, cujos, cujas
Possessives	meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, vosso, vossa, vossos, vossas
Demonstrative	este, esta, estes, estas, isto, esse, esses, essa, essas, isso, aquele, aquela, aqueles, aquelas, aquilo,nessa, desta, daquela, cujo, cuja,cujos, cujas
Subjective Personal	eu, tu, ele, nós, vós, eles, me, te, se, lhe, o, a, nos, vos, lhes, os, as, mim, comigo, conosco, ti, contigo, convosco, si, consigo
Objective Personal	você, vocês, senhor, senhores, senhora, senhoras, senhorita, senhoritas, vossa senhoria, vossas senhorias

We have used both the weighted-sum approach and NSGA to generate the potential solutions. The results achieved by the former presented a premature convergence to a specific region of the search space instead of maintaining a diverse population. Hence, after several trials we did not succeed in finding the Pareto-optimal front but rather than an approximation of the Pareto-optimal solutions. This kind of behavior can be explained by the sensitivity towards weight presented by the weighted-sum approach. In our experiments, we have defined the same weights for both objectives which lead the algorithm to converge to a region with similar error rate but using twice the features. Figure 3 shows the evolution of the population in the objective plane using weighted-sum approach. As we can see from Figure 3, the best solution in this case has more than 100 features.

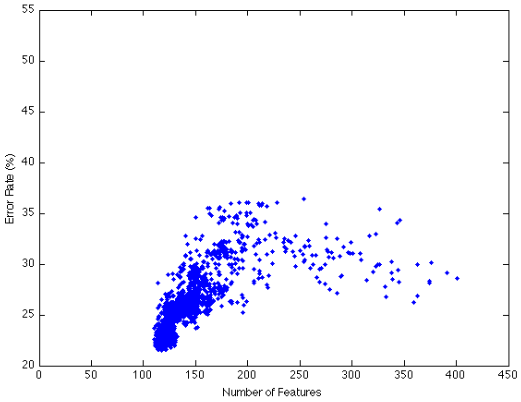


Fig. 3. The evolution of the population in the objective plane using weighted-sum approach (Error rate computed using the searching database).

As we have discussed in Section III-A the Pareto-based approach was designed to overcome this kind of problem. Since NSGA preserves the diversity in the population, this algorithm is able to deal with the problem of converging prematurely to a specific region of the search space. Therefore, it can guide the search towards the Pareto-optimal set. Figure 4 depicts the evolution of the population in the objectives plane from the first generation to the last one. This plot demonstrates the efficacy of NSGA II in converging close to the Pareto-optimal front with a wide variety of solutions.

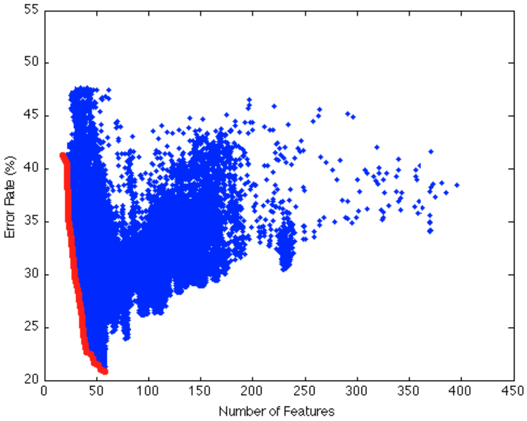


Fig. 4. The evolution of the population in the objective plane using NSGA (Error rate computed using the searching database).

From Figure 4 it is easy to see that several features are correlated or even not relevant as the algorithm converges to a region of the search space with about 50 features. The Pareto-front (marked in red) contains solutions ranging from 25 to 58 features and error rates ranging from 42% to 20%. As discussed previously, after finding the Pareto-optimal front the next step is to choose a solution. In order to perform this task we have used the aforementioned validation database, i.e., all the solutions of the Pareto were test on the validation set and the one that minimized the error was selected. In our experiments, this solution had 58 features.

Thereafter, we trained a new classifier using such a solution using the same databases presented in section III-A. The recognition rate achieved by this new classifier was 74.1% on the testing set. As we can verify, the optimized classifier produced an error rate considerably lower than the original classifier (26% against 42%) with only 58 features from the 408 available ones. This confirms the efficiency of the proposed methodology in selecting a powerful subset of syntactic attributes.

For the problem of feature selection we can observe that the main advantage of the Pareto-based approach is the ability of dealing with different databases with no need of dealing with problems such as scaling and finding the suitable values for the weight vector. Moreover, Pareto-based approaches have the ability of finding the Pareto-optimal front in the first run of the algorithm.

Besides reducing the error rate of the classifier, it is also important to analyze which were the features selected by the genetic algorithm so that we can have a better understanding about the problem. Such an analysis can be useful to help improving and defining new features for the problem of authorship attribution for the Portuguese language. Table V reports the selected features.

TABLE V

SELECTED FEATURES

Group	Quantity	Features
Adverbs	22	lá, dentro, adiante, em cima, ao lado, depois, sempre, com certeza, sem dúvida, ainda, quase, apena, mais, todo, toda, bastante, nada, ninguém, nenhum, antes, qualquer, outro.
Conjunctions	11	porém, por isso, assim como, que nem, segundo, embora, portanto, tais como, contanto que, de modo que, caso.
Pronouns	10	seu, sua, quem , cujo, este, esta, o, a, aquele, onde.
Verbs	15	ser, ver, pular, estar, ligar, estando, efetuando, fazendo, tendo, sendo usando, pagando, aberto, visto, usado

From these experiments we can see that the attributes based on adverbs were selected more often by the search algorithm. On the other hand, some sub-groups of features were not selected at all, e.g., Conjunctions (coordinating additive, coordinating explicative, subordinating final, subordinating proportional), Pronouns (personal), Adverbs (negative).

Since we have 100 authors in the database, analyzing the confusion matrix would be complicated. However, we could get some insight about the problem by analyzing the confusion matrix grouped by subject. Such a matrix can be visualized in Table VI and it shows that the recognition rate in terms of subjects is about 86%. This allows us to split the total error rate of 25.9% into within-class error (14.1%) and between-class error (14%). As expected the class Miscellaneous presents the lowest performance since it gets confused with several other classes. Related classes such as Politics and Economics also feature a high degree of confusion. In such cases the use of subject-dependent words in the feature set could help reducing the confusions.

TABLE VI

CONFUSION MATRIX BY SUBJECTS IN %.

	a. Misc	b. Law	c. Economics	d. Sports	e. Gastronomy	f. Literature	g. Politics	h. Health	i. Technology	j. Tourism
a.	82	7	1	2	1	2	1	2		1
b.	5	84	3		1	2	2	1		
c.	3	3	84			1	4	2		
d.	2		1	86	1	1	1	7	1	
e.		1		1	87	2	1	3		3
f.	4	3	2	1		87	3			
g.	1	1	6			3	88			
h.	1		2	4	3		1	88		2
i.	3	3	3			1	1		89	1
j.	1	1	2		6	1				89

Table VII reports some works on authorship attribution published in the literature. Comparing different works is not a straightforward task since most of the works use different databases and classifiers. However, analyzing Table VII we can see that the results achieved in this work compare to the state of the art.

TABLE VII

PUBLISHED WORKS ON AUTHORSHIP ATTRIBUTION

Ref	Classifier	Database	Rec. Rate(%)
[12]	SVM	web pages	66-80
[5]	SVM	German newspaper	80
[6]	SVM	3 sister's letters	75
[13]	kNN	Novels	66-76
[2]	Distance	Brazilian Novels	78
[7]	SVM	Brazilian Newspaper	72
[3]	Bayes	Mexican poems	60-80
[11]	Bayes	Turkish newspaper	80

VII. CONCLUSION

In this work we have presented a methodology based on a multi-objective genetic algorithm and SVM classifier to select the most discriminative subset of syntactic attributes for authorship attribution. Experiments on a database composed of 3000 short articles written in Portuguese show that the proposed approach is able to find a compact feature that is able to increase the recognition rate in about 15 percentage points. By analyzing the confusion matrix grouped by subjects we could observe that the recognition rate goes to 86%. We believe that this performance can be further improved by adding some subject-related words into the feature set. This is our focus for future works.

REFERENCES

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

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

[3] R. M. Coyotl-Morales, L. Villasenor-Pineda, M. Montes y Gomez, and P. Rosso. Authorship attribution using word sequences. In *Iberoamerican Congress on Pattern Recognition*, pages 844–853, 2006.

[4] K. Deb, A. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transaction on Evolutionary Computation*, 6(2):181–197, 2002.

[5] J. Diederich, J. Kindermann, J. Leopods, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 1, 2003.

[6] M. Gamon. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *20th International Conference on Computational Linguistics*, pages 611–617, 2004.

[7] D. Pavelec, L. S. Oliveira, E. Justino, and L. V. Batista. Using conjunctions and adverbs for author verification. *Journal of Universal Computer Science*, 14(8):2976–2981, 2008.

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

[9] 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.

[10] 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.

[11] T. Tas and A. K. Gorur. Author identification for turkish texts. *Journal of Arts and Science*, pages 151–160, 2007.

[12] Y. Tsuboi and Y. Matsumoto. Authorship identification for heterogeneous documents. *IPSJ SIG Notes*, pages 17–24, 2002.

[13] O. Uzuner and B. Katz. A comparative study of language models for book and author recognition. In *2nd International Joint Conference on Natural Language Processing*, pages 969–980, 2005.

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