Automatic Distinction of Fernando Pessoas' Heteronyms

João F. Teixeira[1] and Marco Couto[2]

[1] University of Porto
[2] HASLab / INESC TEC, University of Minho, Portugal
{jpfteixeira.eng, marcocouto90}@gmail.com

Abstract. Text Mining has opened a vast array of possibilities concerning automatic information retrieval from large amounts of text documents. A variety of themes and types of documents can be easily analyzed. More complex features such as those used in Forensic Linguistics can gather deeper understanding from the documents, making possible performing difficult tasks such as author identification. In this work we explore the capabilities of simpler Text Mining approaches to author identification of unstructured documents, in particular the ability to distinguish poetic works from two of Fernando Pessoas' heteronyms: Álvaro de Campos and Ricardo Reis. Several processing options were tested and accuracies of 97% were reached, which encourage further developments.

Keywords: Authorship Classification, Machine Learning, SVM, Text Mining

1 Introduction

With the dawn of Text Mining (TM) a massive amount of information was enabled to be retrieved automatically. It is intended to find and quantify even subtle correlations over a large amount a data. This way, a wide variety of themes (economics, sports, etc) with different levels of structure, could be analyzed with little effort. Many TM solutions have been employed in security and web text analysis (blogs, news, etc). TM has been used in sentiment analysis as for evaluating movie reviews to estimate acceptability [1]. Forensic Linguistics enhances TM by considering higher level features of text. Linguistic techniques are usually applied to legal and criminal contexts for problems such as document authorship, analysis and measure of content and intent.

The purpose of this study is to generate a small representation of a large corpus of poems, able to discern between authors or aliases. For this initial study we selected to classify the collection of poems by two of Fernando Pessoa's heteronyms. Ricardo Reis and Álvaro de Campos were chosen due to their contrasting themes and initial concerns relative to the model's accuracy for this kind of dataset.

To the best of our knowledge, there are no pattern recognition studies for alias distinction on poetic texts, therefore, no direct comparison of this work can be made. On the other hand, there is research on generic alias identification [2], however the objective is to find which aliases correspond to the same author and not to distinguish between personnas.

The author whose works we analyze is Fernando Pessoa [3], who wrote under several heteronyms or aliases. Each one had their own life stories and personal taste in writing style and theme.

Ricardo Reis is an identity of classical roots, when considering his poems' structure, theatricality and entities mentioned (ancient Greek and Roman references). He is fixated with death and avoids sorrow by trying to disassociate himself with anything in life. He seeks resignation and intellectual happiness.

Álvaro de Campos has a different personality, even presenting an internal evolution. Initially, he is shown to be a thrill seeker, mechanic enthusiast, and wishes to live the future. In the end, he feels defeated by time and devoided of the will to experience life. Consequently, he uses a considerable amount of interjections, in a weakly formatted writing style, with expressive punctuation.

The remaining of this document is structured as follows: In Section 2 the dataset is presented and described. Section 3 shows the methodology employed. Section 4 details the experimental approach, along with the result discussion. Section 5 the overall findings are presented along with possible future work.

2 Dataset

The dataset used in this work consists of the complete known poetic works[3] of *Ricardo Reis* (class RR) and *Álvaro de Campos* (class AC). Table 1 presents some statistics concerning the dataset.

Table 1. Class distribution

			# of Words				# of Verses			
Class	**# of entries**	**%**	**Avg.**	**Std.**	**Min**	**Max**	**Avg.**	**Std.**	**Min**	**Max**
RR	129	54%	77.9	65.1	19	570	14	12	4	106
AC	108	46%	360.9	904.2	29	7857	46	103	5	909

3 Methodology

In this section, the steps taken and experimental approach followed are shown. First, we tested the classifier with the tokenized documents and we progressively introduced other pre-processing models, comparing their performance.

The SVM model was validated with 70% of the dataset using 5-fold cross-validation while the remaining 30% enabled to evaluate the generalization performance of the generated model, i.e., the voting result of the 5 fold models.

[3] Available at: `http://www.dominiopublico.gov.br`

3.1 Document Pre-processing

S_1*-Tokenization.* Each document was turned into a sequence of word-level terms. Then, they were compacted into *bags-of-words* (BoW), disregarding their order, which is the most common document representation [4].

S_2*-Casing Transformation.* After the tokenization, all words suffer a lowercase transformation, reducing the number of different terms. Here, we disregard the capitalization of the poems' first word at every verse, while inadvertently removing significance from capitalized names and some metaphoric references.

S_3*-Length Filtering.* We remove from the token bags terms that contain below 4 or above 15 letters. The reason for this relates to the high probability of shorter words being irrelevant articles or connectors, leading to overfitting of the model, and not many Portuguese words have such large lengths. In fact, the average size of Portuguese words is 4.64 [5]. Nevertheless, removing words produces a more compact representation of the dataset and reduces possible dimensionality issues the classification model may experience with larger feature spaces.

S_4*-Stemming.* Word Stemming also compacts document instances. This consists of removing word affixes, leaving only the root term. Generally, stemmers follow iterative replacement rules, some even dealing with irregular and rare terms. For this work, the Snowball Portuguese dictionary was used [6] [7].

S_5*-Stopword.* Finally, we include stopword removal. This consists of ignoring all terms in a given dictionary. This might help the classifier focus on meaningful terms instead of considering articles, connectors and overall writing style. Also, specific unwanted words can be eliminated.

3.2 Occurrence Metrics

To evaluate if a word is distinctive for the classification task some methods based on its occurrence can be used. The following metrics were experimented:

Binary Term Occurrence. BTO identifies the number of documents in which a given term occurs. It provides little information thus is rarely used.

Term Occurrence. The TO metric provides the number of times a word occurs on each document of the collection. This can be viewed as a measure of significance of a given word for each document.

Term Frequency. The TF is a relative measure of the word occurrence considering the number of words in a document. Consequently, this can be misleading depending on the documents' length variability.

TF-IDF. is generally calculated as the product of TF and the Inverse Document Frequency (IDF) [8]. The IDF approach concerns the number of documents which contain a given term. A term that occurs frequently does not provide discriminative power and should be given less importance (lower weight) [9].

We used SVMs [10] that can linearly separate clusters of data on feature space, by maximizing the hyperspace boundary margin. The model was fed an array of occurrence metrics for the terms included after pre-processing.

Since the focus of this work was on Text Mining, the SVM model employed was relatively simple. A linear kernel with shrinking heuristics was used. It included a termination tolerance $\varepsilon = 0.001$ and no penalty ($C = 0$).

4 Experimental Results

4.1 Estimation using Cross-validation

In Section 3, we conduct several experiments in which the text pre-processing algorithms are incrementally included. These experiments considered the occurrence metric *tf-idf* since it is intuitively the most appropriate to compare the results of models trained with such different instance content. The accuracy of the experiments is S_1:93.35%, S_2:91.58%, S_3:90.44%, S_4:91.03% and S_5:90.44%.

With the length and stopword filtering several of the top scoring terms (SVM weights) were removed. However most of these were articles and connectors which could lead to model overfitting. Their removal only decreased slightly the accuracy. Along with stemming and the lowercase transformation, the number of attributes considered was reduced in more than half (8941 to 4398 terms).

The following experiments aim to evaluate the influence of different the occurrence metrics on the classification model. Table 2 presents the results of those experiments. The results show that the model using Term Occurrence based metrics (BTO, TO) performs worse than with frequency based metrics (TF, *tf-idf*) including misclassification rate balance.

We note that this comparison is not truly fair. The processing pipeline for the experiments was previously optimized for *tf-idf* thus, providing only a general comparison. The performances with these last two are very close and, therefore, the best method cannot be directly found.

Table 2. Occurrence Metrics Experiments (%)

Binary TO			TO			TF			TF-IDF		
Acc	$F1_{RR}$	$F1_{AC}$	Acc	$F1_{RR}$	$F1_{AC}$	Acc	$F1_{RR}$	$F1_{AC}$	Acc	$F1_{RR}$	$F1_{AC}$
80.74	84.91	73.33	68.71	77.39	49.01	91.03	91.89	89.80	90.44	91.58	88.73

4.2 Evaluation of Validation Setup

In this section, we considered, from the previous experiments, the pipelines with the two best performances and with BTO (baseline), while using the remaining 30% of the dataset. The results are shown on Table 3.

As expected, BTO maintained the low accuracy and obtained lower $F1_{AC}$. The best two models kept the high accuracy, however, *tf-idf* managed to overcome the improvement of TF, from the validation phase, even if only by 3 instances. This suggests that, even though *tf-idf* presented lower validation results, it was somewhat underfitting. Either way, these comparative results were expected due to the consideration of term rarity metrics of *idf.*

4.3 Influence of Long Poems

Due to a large difference of length statistics between the two labels, we conducted further analysis. The documents were segmented into multiple instances such

Table 3. Testing Set Results (%)

Binary TO			TF			TF-IDF		
Acc	$F1_{RR}$	$F1_{AC}$	Acc	$F1_{RR}$	$F1_{AC}$	Acc	$F1_{RR}$	$F1_{AC}$
66,20	76,47	40,00	92,96	93,83	91,80	97,18	97,44	96,88

Table 4. Updated Class distribution

			# of Words				# of Verses			
Class	**# of entries**	**%**	**Avg.**	**Std.**	**Min**	**Max**	**Avg.**	**Std.**	**Min**	**Max**
RR	169	51%	63.5	27.0	3	161	11	4	1	17
AC	165	49%	240.8	150.0	29	521	30	145	5	54

that the portions had the maximum amount of verses equal to the previous mean for that class (plus a tolerance). Table 4 presents the updated class distribution.

For this experiment, the complete word processing pipeline and *tf-idf* scoring criteria were used (testing phase best results). For the cross-validation and testing steps, respectively, the accuracy was 96.58% and 96.04%; the $F1_{RR}$ was 96,64% and 96,15% and the $F1_{AC}$ was 96,49% and 95,92%.

The results show that imposing the upper bound on the number of verses per poem increased the accuracy of the model in the validation phase by around 5%. Apart from accuracy, the model should have really improved since the misclassification rates became more balanced.

It is safe to assume that longer poems might be more difficult to sort correctly into classes since they encompass more terms that can be highly influential to the *tf-idf* metric (through emphatic repetition, etc) which may not contribute positively for the accurate learning of attribute weights. Thus, this can affect poorly on the classification. On the other hand, the test results, in a way, contradict the analysis from the cross-validation. It performs slightly worse than the test experiment for *tf-idf.* As of this, we cannot provide an acceptable hypothesis as to which this occurs, rendering this analysis inconclusive.

5 Discussion and Conclusions

In this work we aimed to distinguish the authorship of poetic texts from two heteronyms of Fernando Pessoa, solely using basic Text Mining approaches. To our surprise, the methods were able to predict quite accurately (most over 70%, best ∼97%), further verifying the a clear difference between the heteronyms.

This comes as a revelation mainly because, the author is, in fact, the same, despite having created these two personnas, and thus, the vocabulary and certain parts of writing style should be ubiquitous to the heteronyms.

Many of the best discerning words were related to writing style, including several possessive related terms for AC. However, obviously, some of the best terms were theme related keywords such as *grande* and *sentir*, referring to the magnificence of feelings of AC, *cansaço, domingo* and *sonho* to the tiredness AC feels towards the end and recollections of the past; while RR tries to remain forever calm and avoids pain.

Among the settings tested, *tf-idf* demonstrated, as expected, the best balance and generated the highest accuracy for the testing set.

The change in accuracy for the shorter instance set was not conclusive. These results suggest that dividing the larger poems was either not that relevant or additional instances would be needed to confirm (accuracy already close to 100%).

Although our methodology produces good results, we intend to extend the study to the rest of the heteronyms to evaluate if this kind of simple analysis is still sufficient for discernibility. Additional relevant experiments and approaches could include the model's response to a few verses instead of large chunks or complete poems and word *n-grams* analysis for style traits identification.

We realized that, according to Zipf's law, both the highest and lowest frequent terms are less frequent in large documents. Thus, our approach could be improved concerning the enhancement of term relevance instead of only minding to frequency. This could be done by including a normalization term in the *tf-idf* formula [11].

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 79–86
2. Nirkhi, S., Dharaskar, R.V.: Comparative study of authorship identification techniques for cyber forensics analysis. CoRR **abs/1401.6118** (2014)
3. de Castro, M.G., ed.: Fernando Pessoa's Modernity Without Frontiers: Influences, Dialogues, Responses. Tamesis Books, Woodbridge, Suffolk, UK (2013)
4. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Comput. Surv. **34**(1) (March 2002) 1–47
5. Quaresma, P., Pinho, A.: Análise de Frequências da Língua Portuguesa. In: Livro de Actas da Conferência Ibero-Americana InterTIC, Porto, Portugal, IASK (2007) 267–272
6. Porter, M.F.: Snowball: A language for stemming algorithms. `http://snowball.tartarus.org/texts/introduction.html` (10 2001) [Online].
7. Porter, M.F.: Snowball: Portuguese stemming algorithm. `http://snowball.tartarus.org/algorithms/portuguese/stemmer.html` [Online].
8. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Inf. Process. Manage. **24**(5) (August 1988) 513–523
9. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation **60**(5) (October 2004) 503–520
10. Vapnik, V.N.: An overview of statistical learning theory. Trans. Neur. Netw. **10**(5) (September 1999) 988–999
11. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '96, New York, NY, USA, ACM (1996) 21–29