Detecting Quantum Mysticism in Books Published in Brazil with Data Science

Renato P. dos Santos

12 July 2017

Abstract

Brazilian bookstores have been flooded with titles including the word 'quantum,' but more of 'quantum mysticism' flavour, which purports the existence of links between quantum mechanics and Eastern mysticism, leading to serious misunderstandings. This work aims to identify terms that can help readers, especially high-school teachers, to recognise to which of those categories a book pertains even before reading it through, contributing to breaking the vicious circle of students learning pseudoscience and passing it on as truth. 22 terms were identified that discriminate with an accuracy of 94% between the categories 'quantum mysticism' and science or science popularisation.

Data

The online catalogues of the four largest bookstores in Brazil, Cultura, Saraiva, Amazon and FNAC, were searched in two moments: the first, between March and April 2016, and the second in October 2016. To avoid subjectivity in the selection of Books, objective criteria were defined: in the initial collection of the books, only books containing the words 'quantum' or 'quantum' in their titles were selected; in the second, were included not only books with the words 'quantum' or 'quantum' in their titles that, for some reason, escaped the first collection, but also books that, even without those words in their titles, contained them in the synopsis.

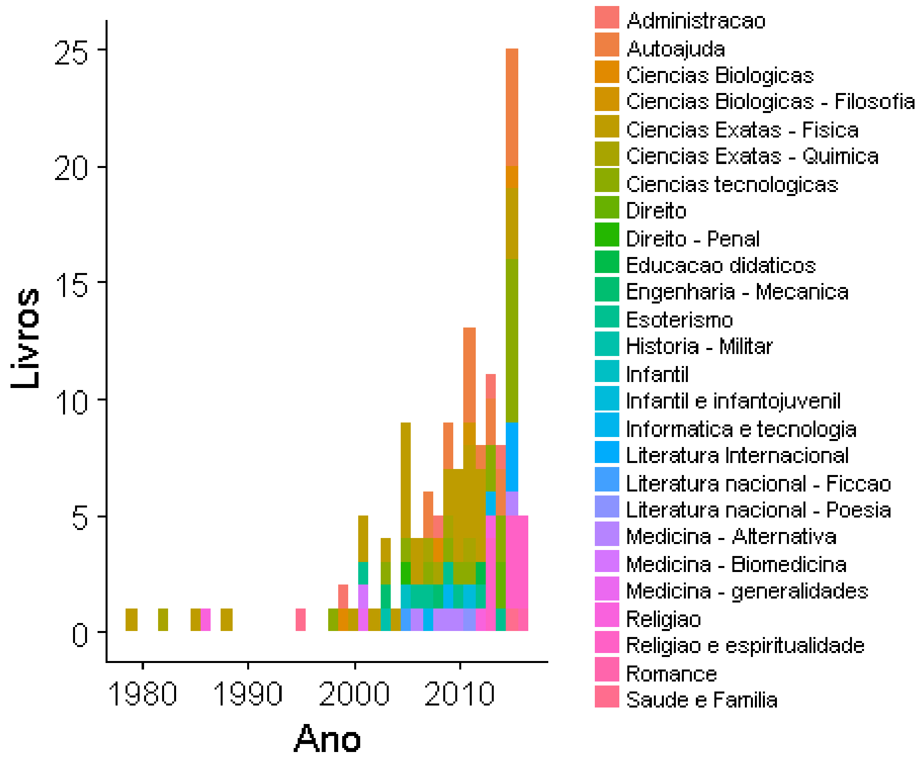
For each book, a record was created containing additional data, such as author's name, synopsis, and publication date. However, only the titles and synopses of the books constituted the corpus of the text analysis. This database was analysed using data science techniques, specifically using the text mining, text analysis, and machine learning from R language (R Core Team, 2016) R version 3.4.1 (2017-06-30). Although R is very versatile and has lots of features, all the work was done in a reasonable time using RStudio, a versatile open-source integrated development environment (IDE) for R, running on a conventional desktop, with x86_64-w64-mingw32/x64 (64-bit) architecture, and Windows 8.1 x64 (build 9600).

The database contains 181 rows (books) and 15 columns (book data).

The 15 variables in the database are: 1 N 2 Titulo 3 Autor 4 Editora 5 Ano 6 Disponibilidade 7 Preco 8 Genero 9 Categoria 10 Link 11 Paginas 12 Origem 13 Nacionalidade 14 Idioma 15 Sinopse

Analysis

The books were originally classified by bookstores in various, and sometimes arbitrary, genres such as 'Self-Help', 'Exact Sciences - Physics', 'Alternative Medicine', or even 'Infantile', as can be seen from the legend in Figure 1. This classification may merit a study by itself, but it would transcend the scope of this work.



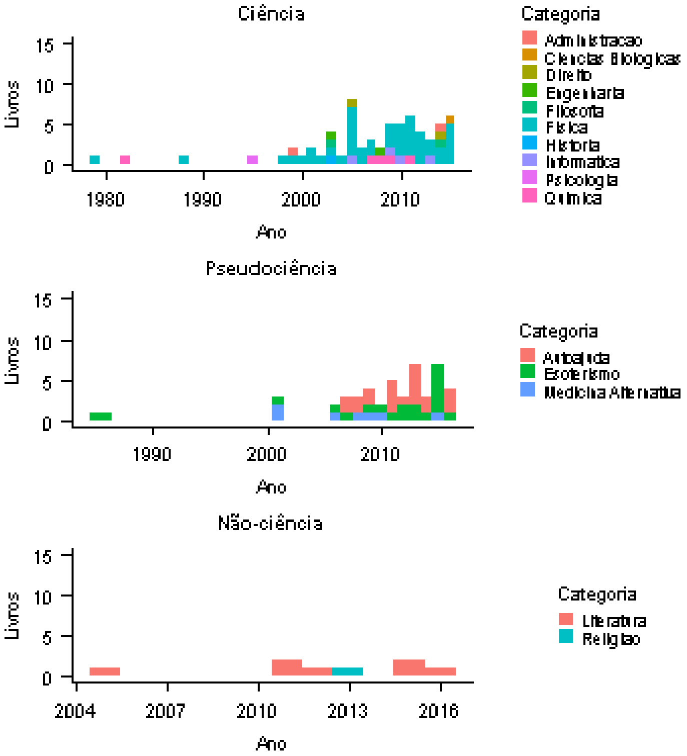
```
## Saving 5 x 4 in image
```

For the purposes of this work, the books were reclassified into only three categories, according to Bunge's (1982) 'demarcation criteria: 'science' (including technology), 'pseudoscience' and 'non-science' (including knowledge fields that are not based on science without, however, coming into conflict with it, such as literature). We did not see the need to include the fourth category proposed by Bunge (1982), 'proto-science', which includes fields of knowledge in the process of becoming scientific. The totals of books in each category are presented in Table 1.

Tabela 1 – Distribuição dos livros segundo as categorias.

GenCat	Freq	%
ciência	91	50
pseudociência	81	45
não-ciência	9	5
Totals	181	100

As the first result of this exploratory analysis, it was verified that the current availability of scientific books dates back to the 1970s, while a growing trend of pseudoscientific works has emerged in Brazil since 2000 (Figure 2).



```
## Saving 5 x 4 in image
```

The 'non-science' category was abandoned because it had little presence in the set of books listed and, mainly, because it does not cause prejudice to the learning in science, since it included only works of literature, such as the science-fiction book 'Quantum Utopy' and 'The mystery of the black sphere and the Quantum box'.

This choice reduced the data set to 168 books.

From the 181 collected books, we selected only those written in Portuguese for the analysis of text, reducing the database to 176.

Using the text mining tm package of the language R (Feinerer; Hornik; Meyer, 2008), the titles and synopses of these books were then transformed into unnoted linguistic corpora for further processing.

The words of the corpora were transformed into lower case letters, punctuation marks and other special characters, excess spaces and numbers. Empty words (*stop words*) were also removed using empty package lists from *package tm*, from Snowball project and ranks.nl.

Instead of constructing indexes (*document-term matrices*), from the corpora, databases were produced in which the words consisted of columns and their relative frequencies in each book in the lines.

Since the syntax of the R language does not allow for accented or special characters, such as a hyphen, in the name of the variables, the words of the titles and the synopses had to be transliterated to their correspondents without an accent and to a point, instead of the hyphen.

To ensure the best discrimination between the 'science' and 'pseudoscience' categories, words associated with both categories simultaneously were discarded.

Due to the small extension of the base, and because the synopses and, even more, the titles of the books, consisted of short texts, the remaining words were sparse, that is to say, with little frequency in the texts, making analysis extremely difficult by usual methods, such as the *bag of words* (Harris, 1954).

Although words with almost constant frequencies in all texts of the corpus, also called predictors of variance close to zero, are generally removed, assuming they are not very informative, this procedure often results in removing the strongest predictors of the model (Gelman et al., 2008) and, therefore, this practice was not followed here. However, words that have constant frequencies in all texts of the corpus have truly zero variance and have been removed.

Another problem arises from highly correlated variables; The usual PCA approach (Kuhn, 2008) could make it more difficult to interpret the predictors and therefore chose instead to simply remove variables with a correlation between them, as measured by the Pearson correlation coefficient, greater than 80%.

After all of these cleaning procedures, the data set has been reduced to 270 words (columns).

Gelman et al. (2008) suggest that the continuous variables should be centred on an average value of *0* and rescaled to a standard deviation of *0.5* to resemble them to binary variables, which assume only the values *0* and *1*. In this case, however, as if only interested in the presence or absence of words, instead of their frequency, the variables were transformed into binary ones, that is, frequency values greater than or equal to zero were transformed to *presente* and *ausente* respectively.

The usual practice of machine learning recommends randomly subdividing the data set and allocating at least 70% of them to the training of the model, reserving the remaining 30% for validation. The R language has the resources to perform this subdivision automatically, taking into account a balance between the data categories in the two subsets.

In order to ensure reproducibility, a global seed (*42*) has been established for the pseudorandom number generator.

After randomly subdividing the data set into approximately 70% of them for model training and 30% for validation, it resulted in sets of 119 and 49 remarks (books) respectively , For the same `n ncol (titlesData [, - (1: 2)])` columns of the headings and `n ncol (synopsesData [, - (1: 2)]) of the synopses words in both sets.

When the data set is small relative to the number of variables, in order to avoid overfitting and to reduce forecasting errors, when applied outside the sample (*out-of-sample errors*), it is usually performed a cross-validation k-fold (*k-fold cross-validation*) (Seni; Elder, 2010, pp. 26-28) on the training data set. In this case, a 7-fold-cross-validation was performed.

A number of the most popular machine learning algorithms available in the *Caret package* (Kuhn, 2008) were tested, including *Support Vector Machines with Radial Basis Function Kernel* (Karatzoglou et al., 2004) *eXtreme Gradient Boosting* (Chen; He; Benetsy, 2016; Friedman, 2001), *Random Forests* (Breiman, 2001; Liaw; Wiener, 2002), and *Recursive Partitioning and Regression Trees* (Breiman et al., 1984; Therenau; Atkinson; Ripley, 2015), among many others.

After applying each model to the training datasets, the Caret package also automatically calculates the usual Jordan accuracy and Cohen Kappa index parameters to evaluate the expectations of the models' ability to classify the test data. The results of these parameters for titles and synopses are presented in Tables 2 and 3.

Comparação entre modelos para títulos

ModelNames	Performances
SVM	svmRadial, 0.958, 0.915
XGBoost	xgbTree, 0.529, 0
RForest	rf, 0.605, 0.169
RPART	rpart, 0.63, 0.224

Comparação entre modelos para sinopses

ModelNames	Accuracy	Kappa
SVM	0.983	0.966
XGBoost	0.756	0.495
RForest	0.765	0.513
RPART	0.807	0.6

From these tables, notes that the SVMRadial model is the one that looks most promising, with the RPART second, in the set titles and synopses.

Now, models must be validated by applying them to the test data sets, both title words and synopses, with 49 books each, simulating the actual case of a visitor choosing books in the Bookstores for their titles and synopses. The results can be seen in the confusion matrices of Tables 4, 5, 6, and 7.

Matriz de confusão do modelo SVMRadial para títulos

	pseudociência	ciência	class.error
pseudociência	22	1	0.05
ciência	0	26	0.00

Matriz de confusão do modelo SVMRadial para sinopses

	pseudociência	ciência	class.error
pseudociência	21	2	0.1
ciência	0	26	0.0

Matriz de confusão do modelo RPART para títulos

	pseudociência	ciência	class.error
pseudociência	10	13	1.3
ciência	0	26	0.0

Matriz de confusão do modelo RPART para sinopses

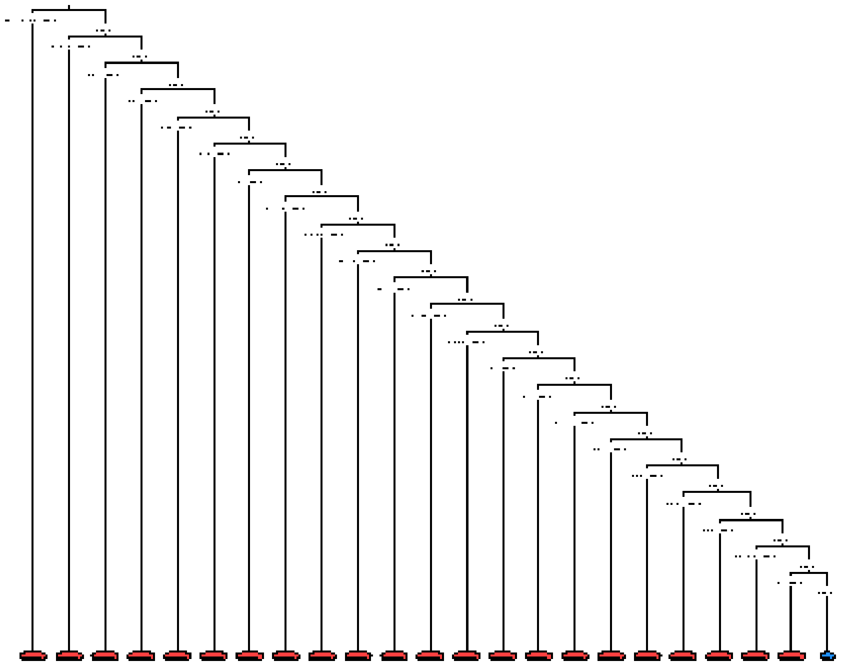
	pseudociência	ciência	class.error
pseudociência	10	13	1.3
ciência	0	26	0.0

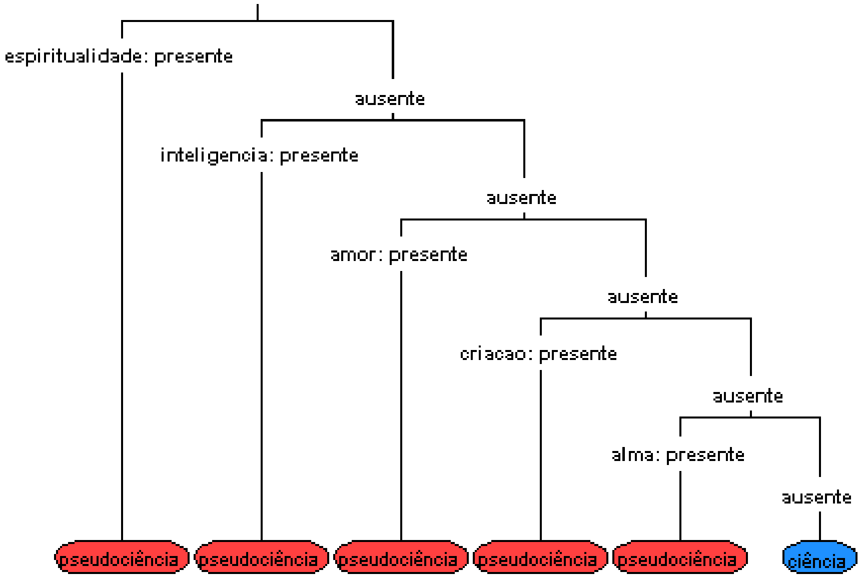
As can be seen from the tables, all 26 test set scientific books were correctly classified in the 'science' category, but not all pseudoscientific books were recognised as By their titles. The RPART model was the one that had the lowest classification error rate (130%), by title words.

The 13 books that were incorrectly classified by their titles are:: A cura Quântica, Quântica - O caminho da felicidade, Sociedade Quântica, A dimensão Quântica da realidade, A realidade quântica: nos confins da nova física, Toque Quântico 2.0, Aumento da Potência do Toque Quântico, O segredo dos mestres e o mundo quântico, Espaço, tempo e espírito - Espiritismo e Física Quântica, Origem, Evolução e Destino da Consciência,da Vida e do Universo: Cosmovisões jônica, quântica, gênica e fractal, Ascensão Profissional - Física Quântica Aplicada Às Relações Humanas, A Força da Calma no Xamanismo de Jorge Menezes , Economia da Consciencia

The model is an economic one, since it consists of a regression tree with only 22 'sheets' (Figure 4a), related to the following 22 title words: espiritualidade.L, inteligencia.L, amor.L, criacao.L, alma.L, ciencia.L, poder.L, psicologia.L, criatividade.L, medicina.L, medico.L, enigma.L, ativista.L, psi.L, fator.L, raciocinio.L, cerebro.L, faraos.L, paradoxo.L, alem.L, visionaria.L, tao.L.

This model discriminates with reasonable success among categories, so that the presence of any of those words in a book synopsis indicates that it belongs to the category "quantum mysticism" and not to "science", as seen from the confusion matrix of Table 2 for the titles and that of Table 3 for the synopses.




espiritualidade: presente
ausente
inteligencia: presente
ausente
amor: presente
ausente
criacao: presente
ausente
alma: presente
ausente
pseudociência
pseudociência
pseudociência
pseudociência
pseudociência
ciência

References

- Breiman, Leo et al. *Classification and Regression Trees*. London: Chapman and Hall/CRC, 1984.
- Breiman, Leo. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001.
- Friedman, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189-1232, Oct. 2001.
- Gelman, Andrew et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, v. 2, n. 4, p. 1360-1383, Dec. 2008.
- Hastie, Trevor J.; Pregibon, Daryl. Generalized linear models. In: Chambers, John M.; Hastie, Trevor J. (Org.). . *Statistical Models in S*. Boca Raton, FL: CRC Press, 1991.
- Kuhn, Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, v. 28, n. 5, Nov. 2008.
- R Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2016. Available at R site: https://www.R-project.org/.
- Seni, Giovanni; Elder, John F. *Ensemble Methods in Data Mining*: Improving Accuracy Through Combining Predictions. San Rafael, CA: Morgan & Claypool, 2010.
- Suykens, J.A.K.; Vandewalle, J.; De Moor, B. Optimal control by least squares support vector machines. *Neural Networks*, v. 14, n. 1, p. 23-35, Jan. 2001.