Quantitative and Qualitative Approach of Scientific Paper Popularity by Naïve Bayes Classifier

Tobias Ribeiro Sombra, Rose Marie Santini, Emerson Cordeiro Morais, Walmir Oliveira Couto, Alex de Jesus Zissou, Pedro Silvestre da Silva Campos, Paulo Cerqueira dos Santos Junior, Glauber Tadaiesky Marques, Otavio Andre Chase, José Felipe Souza de Almeida
Brazilian Institute of Information, Science and Technology (IBICT) - Federal University of Rio de Janeiro (UFRJ), Brazil

ABSTRACT

Usually, scientific research begins with the collection of data in which online social media tools can be some of the most rewarding and informative resources. The extensive measure of accessible information pulls in users from undergraduate students to postdoc. The search for scientific themes has popularized due to the availability of abundant publications that resides in scientific social networks such as Mendeley, ResearchGate etc. Articles are published on these media inform of text for knowledge dissemination, scientific support, research, updates etc, and are frequently uploaded after its publication in a proceedings or journal. In this sense, data collected from database often contains high noise and its analysis can be treated as a characterization undertaking as it groups the introduction of a content into either good or bad. In this text, we present quantitative and qualitative analysis of papers popularity in Mendeley repository by using naive Bayes Classifier.

Keywords: Scientific Social Networks; Mendeley; Naïve Bayes; Machine Learning

1. INTRODUCTION

Machine learning presents several concepts that are similar to each other (Shalev-Shwartz and Ben-David, 2014). As an example, Rocha et al. (2008), who conceptualizes machine learning as the ability of a computer program to learn through classifiers. Among them, it is possible to highlight the Neural Networks, Bayesian Networks, among others, each carrying a specific paradigm and definitions. In addition, Faceli et. al. (2011), says that machine learning is the ability of computational tools to create hypotheses or functions on their own that can solve a certain problem, through experience during automatic learning. Likewise, Mitchell (1997) states that Machine Learning are computational algorithms that aim to automatically improve with experience.

According to Rocha et al. (2008), supervised learning can be exemplified through the presence of a teacher, that is, each example presented in the data set contains a correct answer that would be the exit class. In this case, each example must contain its input attributes and the corresponding output classes. Conduta and Magrin (2010) categorize the output classes into two types: if the classes have discrete values,

the problem is categorized as classification; if the classes have continuous values, they are categorized as regression.

Another concept related to error measures for classification problems is the Confusion Matrix (Rocha et al., 2008), which is based on mapping the examples contained in the training to verify how many were predicted. It consists of a table with distributed values, representing the total of examples trained by a classifier, and a higher concentration of values is expected on its diagonal, that is, where it will always present the negative-negative and positive-positive relationship to identify the values expected. These values are examples that have been correctly classified by a machine learning algorithm. As a consequence, the Percentage of Examples Correctly Classified (PECC), counts the number of examples for which the predicted value for the class coincides with the real value, i.e., diagonal values of the matrix. PECC is usually normalized in terms of percentage, dividing by the total number of examples.

Bayes' theorem may be used in any situation where needs to calculate conditional probabilities after collecting data and is considered one of the simplest yet complete for data classification (Rocha *et al*., 2008). When applied in the form of an algorithm, the probability calculations assumes that the presence of a particular feature in a class are obtained through the co-occurrence frequencies for each attribute of the training data set. As a result, the algorithm concludes the classification based on the likelihood concept, which is obtained by multiplying the relative frequencies of each attribute present in the test example, associating each output class described in the database.

The naive Bayes classifier has been widely used in the scientific community and, as examples, are the works of Carvajal et al. (2015), Xu (2018) and Li et al. (2020). The first uses this algorithm to classify, predict and represent associations between pathogen reduction and operational conditions, a need that arose due to the interest in optimizing risk management during biological treatment processes. The second uses the textual classification algorithm, in order to categorize and provide conceptual visualization of the document collection. The third uses the algorithm as a way to protect data privacy due to the ability to group probability information.

This article is based on the continuity of a formulation presented by Sombra et al. (2020), but within the context of qualitative analysis. Thus, the work consists of three stages, which are: data collection and selection, pre-processing and mining. The first step aims to obtain data on the Mendeley platform by developing an application based on the Mendeley API (https://dev.mendeley.com). The second stage aims to treat the collected data, in order to eliminate existing noise. In this case, treatments were carried out to remove repeated documents. The third step, so-called mining, consists of treating the documents to be executed in the naive Bayes algorithm. The algorithm presents a model based on the UCI Machine Learning Repository database (https://archive.ics.uci.edu/ml/index.php), which is a repository of Machine Learning databases developed by the University of California Irvine that presents standards for data set composition. Another important factor to be mentioned is the work organization, which was divided into subcategories. In this sense, Proceedings, Open_Proceedings, Open_Journal and Journal that present the names five output classes, three output classes and two output classes, as subcategories.

2. QUANTITATIVE APPROACH

Based on the Mendeley API, at no less than 16,091,264 documents were collected using the words proceedings and journal as query words. Subsequently, the final subset was divided, as follows: 1,696,118 for Proceedings, 3,416 for Open_Proceedings, 815,794 for Open_Journal and 13,575,936 for Journal.

In order for the data to be adapted, the discretization process was carried out, which consists of converting the attributes of the documents to nominals, since the naive Bayes algorithm presents the supervised paradigm. The discretization process aims to establish ranges of values according to the attributes and assign a name to each range. The calculation used basically works as a rule of three, being a ratio between the value to be discretized and the highest value in the database. Table 1 shows the frequency distribution for the category Open_Proceedings in the subcategory five output classes.

. Table 1. Distribution of discretization for the category Open_Proceedings in the subcategory five output classes

Output Classes	Discretizations (%)	Discretizations (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% a 1.6%	0 to 1.04	179	1
Little_Popular	>1.6% to 4%	>1.04 to 2.6	77	2
Popular	>4% to 7%	>2.6 to 4.55	60	3.36
Very_Popular	>7% to 12%	>4.55 to 6.5	22	5.54
Extremely_Popular	>12% to 100%	>6.5 to 65	21	17.9

Table 1 shows the average frequency of popularity between the number of readers for each output class and the total number of documents. These values represent the averages corresponding to the concentration of most of the reader counter for each class of output. The discretization column (%) was obtained by means of exhaustive tests in the database until finding an adjustment capable of allowing the distribution balance or a decreasing distribution. Thus, the output class Not_Popular corresponds to the largest number of documents, followed by Little_Popular as the second largest amount, and so on.

Table 2 and Table 3 show the distributions for the Open_Proceedings category in the subcategories three output and two output classes, respectively.

Table 2. Distribution of discretization for the Open_Proceedings category in the subcategory three output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 1.6%	0 to 1.04	179	1
Popular	>1.6% to 6%	>1.04 to 3.09	115	2.33
Extremely_Popular	>6% to 100%	>3.09 to 65	65	9.01

Table 3. Distribution of discretization for the Open_Proceedings category in the subcategory two output classes.

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 4%	0 to 2.6	253	1.3
Extremely_Popular	>4% to 100%	>2.6 to 65	103	6.79

Using the same procedure, Table 4, Table 5 and the Table 6 show the distributions for the Proceedings category in the subcategories five, three and two output classes.

Table 4. Distribution of discretization for the Proceedings category in the subcategory five output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.5%	0 to 10.38	51,795	3.49
Little_Popular	>0.5% to 2%	>10.38 to 41.54	13,647	20.29
Popular	>2% to 5%	>41.54 to 103,.85	328	63.55
Very_Popular	>5% to 8%	>103.85 to 166.16	840	129.19
Extremely_Popular	>8% to 100%	>166.16 to 2,077	605	307.77

Table 5. Distribution of discretization for the Proceedings category in the subcategory three output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.5%	0 to 10,.38	51,795	3.49
Popular	>0.5% to 2%	>10.38 to 41.54	13,647	20.29
Extremely_Popular	>2% to 100%	>4.54 to 2,077	5,173	102.77

Table 6. Distribution of discretization for the Proceedings category in the subcategory two output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.5%	0 to 10..38	51,795	3.49
Extremely_Popular	>0.5% to 100%	>10.38 to 2,077	18,820	42.96

In continuation, Table 7, Table 8 and Table 9 show the distributions for the Open_Journal category in subcategories five, three and two classes of output.

Table 7. Distribution of discretization for the Open_Journal category in the subcategory five output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.035%	0 to 3.2	60,211	1.82
Little_Popular	>0.035% to 0.08%	>3.2 to 7.4	40,069	5.30
Popular	>0.08% to 0.15%	>7.4 to 13.9	28,263	10.11
Very_Popular	>0.15% a 0.3%	>13.9 to 27.9	23,033	18.97
Extremely_Popular	>0.3% a 100%	>27.9 to 9,326	14,184	59.51

Table 8. Distribution of discretization for the Open_Journal category in the subcategory three output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.06%	0 to 5.59	83,448	2.55
Popular	>0.08% to 0.3%	>5.59 to 27.9	68,128	12.20
Extremely_Popular	>0.3% to 100%	>27.9 to 9,326	14,184	59.51

Table 9. Distribution of discretization for the Open_Journal category in the subcategory two output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0,1%	0 to 9.3	112,316	3.77
Extremely_Popular	>0.1% to 100%	>9.3 to 9,326	54,134	27.82

At least, Table 10, Table 11 and Table 12 show the distributions for the Journal category in subcategories five, three and two classes of output.

Table 10. Distribution of discretization for the Journal category in the subcategory five output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.17%	0 to 16,9	1,516,506	2.21
Little_Popular	>0,17% to 0.5%	>16.9 to 49.79	960,451	30.97
Popular	>0.5% to 0,9%	>49.79 to 89.60	433,574	55.82
Very_Popular	>0.9% to 2%	>89.60 to 199.12	328,187	108.53
Extremely_Popular	>2% to 100%	>199.12 to 9,956	112,695	207.18

Table 11. Distribution of discretization for the Journal category in the subcategory three output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.25%	0 to 24..89	1,702,893	2.98
Popular	>0.25% to 0.8%	>24.89 to 1991.2	1,015,640	30.74
Extremely_Popular	>0.8% to 100%	>1991.2 to 9,956	832,880	1993.7

Table 12. Distribution of discretization for the Journal category in the subcategory two output classes

Output Classes	Discretization (%)	Discretization (Amount)	Total of Documents	Average Frequency of Popularity
Not_Popular	0% to 0.5%	0 to 49.78	2,286,075	2.98
Extremely_Popular	>0.5% to 100%	>49.78 to 9,956	1,065,338	55.40

The tables presented show that although the distribution of the output classes are not balanced, they were ordered from the highest to the lowest according to the class of least popularity to the most popular. In this sense, the data configured in this way, reflect the best reality on the databases, in which most documents do not have as many readers. Consequently, the class Not_Popular in all distributions kept in discretization, presented a greater number of documents. In fact, the percentage distribution helped to improve all output classes, since if it were an equal distribution, this class would have the vast majority of documents, at the risk of the other classes presenting very little or even no value. All percentage ranges presented in the tables were obtained based on exhaustive tests until a range was found that keeps the data balanced or with a decreasing order of magnitude of distribution. Therefore, the Not_Popular class obtains more documents than the Little_Popular class, which in turn has more documents than the Popular class, and so on.

3. QUALITATIVE APPROACH

From the discretization process, the transformation of numerical to nominal attributes allowed the recognition of patterns by the naive Bayes algorithm. All processes of discretization, pre-adaptation and final adaptation were done automatically by means of algorithms that were developed for this purpose. Therefore, the results obtained are based on the continuity to the work of Sombra *et. al.* (2020), in which the qualitative analysis of the same problem was carried out.

The quantitative analysis showed that the subcategory two classes of output, presented better result for having higher PECC than the other subcategories. In this analysis, the output classes on the popularity of the documents consider the quantities of accesses. In short, Table 13 presents a brief summary of the quantitative results in order to explain the qualitative analysis later.

Table 13. Test examples classification in the naive Bayes algorithm for each subcategory

Five Output Classes						
-	Not_Popular	Little_Popular	Popular	Very_Popular	Extremely_Popular	PECC
Open_Proceedins	3	1	-	1	-	53%
Proceedings	3	1	-	1	-	74%
Open_Journal	1	1	2	-	1	44%
Journal	1	1	1	2	-	47%
Three Output Classes						
-	Not_Popular	-	Popular	-	Extremely_Popular	PECC
Open_Proceedins	3	-	-	-	-	54%
Proceedings	3	-	-	-	-	75%
Open_Journal	2	-	1	-	-	64%
Journal	1	-	1	-	1	60,5%
Two Output Classes						

-	Not_Popular	-	-	-	Extremely_Popular	PECC
Open_Proceedins	1	-	-	-	1	73%
Proceedings	-	-	-	-	2	77%
Open_Journal	1	-	-	-	1	76%
Journal	1	-	-	-	1	77%
TOTAL	20	4	5	4	7	-

Table 13 shows that the subcategory two output classes produced a better PECC result, in relation to the others. Another interesting detail is that most of the examples were classified as Not_Popular and followed by Extremely_Popular as runner-up. Besides, there was already expected when considering the subcategories three output classes and two output classes present a lower number of classes. If we disregard the classes which do not appear for all subcategories, there are at least 20 examples for Not_Popular and 8 for Extremely_Popular. This last result was also expected, considering the distribution of discretization previously presented, in this text.

To evaluate the attributes related to the Not_Popular and Extremely_Popular classification test examples, it is necessary to understand what these attributes are and how they were categorized. Table 14 presents all the attributes which can be contained in the test examples related to this type of metric.

Table 14. Attributes and possible characteristics for classification test examples

Attributes	Characteristics
Title	very_bad, bad, good, very_good, excellent
Type	Dependent on database requirements
Source	Dependent on database requirements
Year	Until 1999: classic _article From 2000 to 2007: review From 2008 to 2011: citation From 2012 to 2015: state_of_art From 2016 to 2017: current
Keywords	very_bad, bad, good, very_good, excellent
Authors	very_bad, bad, good, very_good, excellent
Month	Full name of the corresponding month of the document.
Abstract	very_bad, bad, good, very_good, excellent
Reader Count	very_popular, little_popular, popular, very_popular, extremely_popular

The attributes described in Table 14 play an important role in generating results to this type of evaluation process. It is worth mentioning that the Reader_Count attribute is the number of readers for each document and which defines the output classes of the database. In general, they were chosen due to the assumption of helpers as indicators, in order to identify the popularity of an article. The possible

characteristics were thought of as a strategy, so that the naive Bayes algorithm has the ability to identify and learn from the data, therefore, they will be used after the entire data processing process. In this sense, it should be noted that the Title, Keywords, Authors, Abstract and Reader_Count attributes were defined based on the discretization performed previously. Thus, the Reader Count attribute is the output class of the naive Bayes algorithm and the underline names assigned to the possible characteristics were defined only as a means for the algorithm to understand and classify the examples. The Month attributes of publication in the database, are ordered from lowest to highest recurrence. Moreover, the Year attribute is validated until 2017, as it was when the data collection had been completed. Type and Source attributes depend on the database, since Mendeley presents several types of documents, as well as places where they were published. Table 15 shows the number of attributes found for the Not_Popular class.

Attributes	Result									
Title	very_bad	8	bad	4	excellent	5	good	3	-	
Type	Journal	17	Conference_ Proceedings	2	Generic	1	-		-	
Year	classic_ article	10	citation	4	state_of_art	2	review	2	current	2
Keywords	very_bad	17	excellent	3	-		-		-	
Authors	very_bad	18	bad	1	good	1	-		-	
Abstract	very_bad	12	bad	3	excellent	4	good	1	-	

Table 15. Quantitative of attributes found in the test examples for the Not Popular class

Table 15 shows the very_bad characteristic for the frequencies of Title, Keywords, Authors and Abstract, which may be one of the justifications for the classification as Not_Popular. This result indicates that the attributes mentioned above have a small score in relation to the database, which is indicated by the aforementioned characteristic. The Type, predominantly in Journal, only indicates that the majority of classified documents are hosted in scientific journals. The Year, predominantly in classic_article, shows that classified documents were until to the 1999 period, which may indicate that articles do not reach popularity status over time. The questions regarding these attributes can be further enriched after analyzing the test examples at extremely_popular, shown in Table 16.

Table 16. Quantitative of attributes found for the test examples in the extremely_popular class

Attributes	Result					
Title	very_bad	5	excellent	2	-	-
Type	Journal	6	Generic	1	-	-
Year	classic_article	2	citation	2	review	1
Keywords	very_bad	6	excellent	1	-	-
Authors	very_bad	7	-	-	-	-
Abstract	very_bad	5	excellent	2	-	-

Table 16 shows the very_bad characteristic for the frequencies of Title, Keywords, Authors and Abstract, as reported for the Not_Popular class. This suggests that the attributes mentioned may not directly influence the popularity of papers. However, Type remained high in Journal, indicating the preference of researchers when publishing articles in scientific journals. The Year brought a balance between classic_article and review, showing that some documents considered popular by the Naive Bayes algorithm were published until 1999 and the interval between 2008 to 2011.

Other tests were done, considering the Source and Month attributes, for the Not_Popular class. In short, two characteristics in Source appeared evidenced, when the terms proceedings_in_indian_academy_of_sciences_chemical_sciences were used, with frequency 5 and proceedings_in_national_academy_of_sciences_usa, with 4 test examples. In the case of the Month attribute, March was the one that most appeared with 4 examples, followed by April and December tied at 3. The use of the Source and Month attributes were made to identify where documents are most frequently published and what are the months of the year with the highest concentration of publications, respectively. However, the data obtained are still not considered sufficient to make a more specific conclusion on these attributes, requiring further research. The Extremely_Popular class for the Source proceedings_in_national_academy_of_sciences_usa appeared with 2 classified test examples and in Month, November and January, which have been tied in 2 examples for each.

5. CONCLUSION

Qualitative and quantitative evaluation of a data set are very useful when approaching common problems in pattern recognition situations from the viewpoint of Naive Bayes classifier. Based on the scientific social networks Mendeley platform, this work presented the continuity of the results of a previous work, which deals with qualitative analysis with a focus on the popularity metric for technical-scientific publications. Therefore, we present two corresponding situations on the same problem: the percentage discretization model taking into account the PECC and the frequency of access to the attributes which determine the classes related to the metrics of popularity for papers. In this sense, the correlation between qualitative and quantitative analyze allows the assessment of the characteristics of a data set in a way which naive Bayes classifier is suitable for this type of approach. In summary, the optimal selection of particular

pattern classes for identification may thus be approached initially at the level of data qualitative analysis before embarking upon more complex issues of quantitative evaluation.

REFERENCE

CARVAJAL, G.; ROSER, D. J.; SISSON. S. A.; KEEGAN, A.; and KHAN, S. J. Modelling Pathogen $\log_{10}$ Reduction Values Achieved by Activated Sludge Treatment Using Naïve and Semi Naïve Bayes Network Models. **Water Research**, v. 85, p. 304-315, nov. 2015.

CONDUTA, B.; MAGRIN, D. Machine learning. Federal University of Campinas, Limeira, 2010.

FACELI, K; LORENA, A. C; GAMA, J; CARVALHO, A. C. P. L. F. **Artificial Intelligence: A Machine Learning Approach**. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 2008.

LI, L.; WU, W. and XUE. D. Transfer Naive Bayes Algorithm with Group Probabilities. **Applied Intelligence**. v. 50, n. 1, jan. 2020.

MITCHELL, T. M. **Machine Learning**. McGraw-Hill, USA, 1997.

ROCHA, M., CORTEZ, P. & Neves, J. **Intelligent Data Analysis - Algorithms and Implementation in Java**. Lisboa: FCA – Editora de Informática, 2008.

SHALEV-SHWARTZ, S. and BEN-DAVID S. **Understanding Machine Learning: From Theory to Algorithms**. Cambridge University Press, UK, 2014.

XU, S. Bayesian Naive Bayes Classifiers to Text Classification. **Journal of Information Science**. v. 44. n. 1. p. 48-59, fev. 2018.