Identifying the Main Problems in IT Auditing: a Comparison Between Unsupervised and Supervised Learning

Patrícia Maia[1], Leonardo Sales[1], and Rommel N. Carvalho[1,2]

[1] Department of Research and Strategic Information
Brazilian Office of the Comptroller General
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro
Brasília, Distrito Federal, Brazil
{patricia.maia,leonardo.sales,rommel.carvalho}@cgu.gov.br
http://www.cgu.gov.br
[2] Department of Computer Science
University of Brasília
Campus Universitário Darcy Ribeiro
Brasília, Distrito Federal, Brazil
http://www.cic.unb.br

Abstract. One of the main challenges faced by the Brazilian Office of the Comptroller General (CGU) is applying consistent knowledge discovery tools and methodologies to learn from several years of auditing experience from hundreds of thousands of auditing reports with millions of pages it produced during these years. More specifically, we tackle the problem of identifying the most common topics in a context of Information Technology audits performed in Brazil since 2011. In order to tackle this problem, we compare two different approaches, supervised and unsupervised learning. On the one hand, the supervised learning approach generated a model that achieved around 73% accuracy for seven categories using random forest. On the other hand, the unsupervised learning approach using Latent Dirichlet Allocation (LDA) generated a model with five topics, which was considered the best model based on the validation performed by the subject matter experts (SME) from CGU. Nevertheless, it is important to note that both approaches, although implemented independently, generated very similar topics. This also reinforces the success in identifying the main problems found during all these years of IT auditing at CGU using consistent and well-known knowledge discovery methods.

Keywords: LDA, text mining, classification, auditing, IT, topic modeling

1 Introduction

The Brazilian Office of the Comptroller General (CGU) is an agency responsible for, among other things, auditing all contracts and spending related to the federal budget from the Executive branch of the Brazilian Government.

One of the main challenges faced by the CGU is applying consistent knowledge discovery tools and methodologies to learn from several years of auditing experience from hundreds of thousands of auditing reports with millions of pages it produced during these years.

The Audit and Inspection area is responsible for carrying out audits and inspection activities to check how public resources are being used. This task is carried out by CGU through the Federal Internal Control Secretariat, which is the unit in charge of evaluating the implementation of Federal Government budgets, inspecting the implementation of governmental programs, and auditing the management of federal public and private agencies and organizations, among other functions [13].

The result of conducted audits by CGU are stored in a report with all information about the scope, minister, program, findings, problems discovered, among others. This kind of report can be consulted in the Novo Ativa System. This system has all information about the audits conducted by CGU. However, all texts and information are stored without a well-defined categorization.

It is important to be able to retrieve more information from this data. For instance, what are the most common problems in the audits, or if the problems are different among regions of the country. Beyond that, the SFC wants to define a categorization for all kinds of problems found in reports and put this information in each new report that will be included in the system.

We try to solve the problem using two different approaches in parallel. One of them consisted in using a topic modeling technique to discover the main topics using LDA. The other approach used was classification. A small part of the data was manually classified and used to train a model in order to apply it to the remaining data. This paper will discuss both of them and compare the results.

The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 describes our approach, divided in supervised and unsupervised learning. Finally, Section 4 presents the conclusion.

2 Related Work

Debasis [9] works with StackOverflow "linked questions" retrieval. Manual annotations, *e.g.*, tags and links, of user generated content in community question answering forums and social media play an important role in making the content searchable. This work tries to reduce the manual effort by automatizing the search process to suggest a list of candidate documents to be linked to the new document, using topic models. The experiment shows that topic distributions results in a significant improvement in retrieval of the candidate set of related documents.

David [2] considered the problem of a user navigating an unfamiliar corpus of text documents where document metadata is limited or unavailable, the domain is specialized, and the user base is small. Their work proposed to augment standard keyword search with user feedback on latent topics. These topics are automatically learned from the corpus in an unsupervised manner and the users feedback is used to reformulate the original query. The model gives users the ability to provide feedback at the latent topic level.

Shinjee [16] proposed a unified topic model employs two LDA models, one for similar TV user grouping and the other for TV program recommendation. The unified model identifies the semantic relation between TV user groups and TV program description word groups more meaningful that the TV program recommendations usually made. Beyond that, the new model allows users with similar tastes can be grouped by topics and recommended as social communities. The unified model can make recommendations with results 6.5% betters than just use topic models for TV users.

Sales [18] presents a supervised learning model to prevent default risk in public contracts in the Brazilian Government using Logistic Regression and Decision Trees Algorithms. Besides various databases related to public contracts in Brazil, like registry information of hired companies and operational capacity indicators, the model also used CGU audit findings to predict the “bad companies”.This supervised model achieved an average accuracy of 64%.

3 Experiment

The data available in the Novo Ativa System consists of approximately three hundred thousand findings. Finding is a significant fact reported by the designated civil servant during the auditing proceeding. Each of these findings has an key, year, resumed text, detailed text, kind of resource, government program, and location. The data was separated by kind of resource and we used IT resources for this research. We analyzed 2,500 findings related to different IT audits. The detail text range for these findings varied from 1 to 20 pages. The resumed text for each finding is around 2 or 3 lines.

Although we focused in IT audits, in future works we will apply the same methodology in other contexts, such as: education; health; among others.

The data was load in RStudio for pre-processing. In this stage we applied techniques for removing stopwords and low frequency words, changing upper to lower case letters, removing punctuation, accentuation, numbers, and white spaces. These techniques are common during the text mining process and are discussed in more detail in [8,10,3].

For implementing the techniques of text mining, we used the text mining framework provided by the TM package in R studio [3]. This package makes it possible to process, organize, transform, and analyze textual data.

The term-document matrix (TDM) was constructed using the Term Frequency - Inverse Document Frequency (TF-IDF) parameter. TDM is a matrix where the rows represent the words (terms) and the columns represent the documents. For more details in TDM and TF-IDF, see [4,14]. Several tests were implemented in this data using unigrams, bigrams, and trigrams in TDM. The bigrams and trigrams presented the best results [19].

[3] http://cran.r-project.org/web/packages/tm/index.html

After some tests, with the help of the SME, we noticed that some words that remained in the model were not helpful in identifying the latent topic. Therefore, these words were included in the stopwords list.

3.1 Unsupervised Learning

After all the pre-processing, we work with an unsupervised learning, using LDA [4,20]. LDA is a topic modeling technique that consists in defining the most relevant terms in a topic, based in the distance and the proximity of the terms, given a number of topics. This technique calculates the probabilities of one document belonging to any of the topics. The topic with the highest probability will be the one in which the document will be designated.

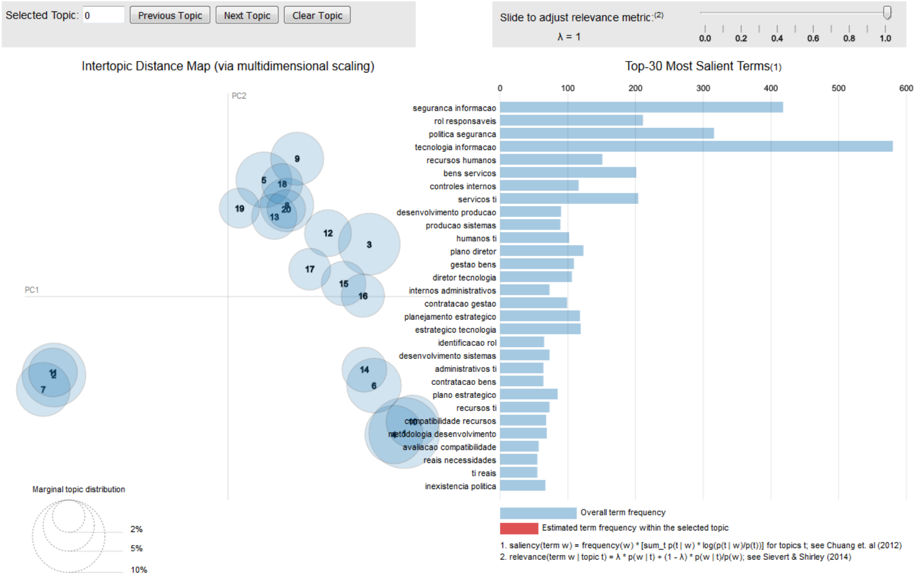
First we defined 20 topics but some of them are very close, including overlaying sometimes. Then we tested several number of topics (20, 15, 12, 10, 7, 6, 5, and 4) and the best results were 5 topics. The results were demonstrated using the LDAvis[4], an R package for LDA visualization. In this kind of visualization, it is possible to see the most frequent topics, the proportion of topics in the set, and the proportion of the terms existing in a topic compared to the others. As shown in Figure 1, each circle defines a topic and the size of this topic defines the proportion of this topic among all existing findings. The blue bars designs the 30 most frequent terms in all topics. The red bars show the most frequent terms in a specific topic. This plot was generated using 20 topics. As it is possible see, this number is too big for this context because the circles are very close or overlaying. This fact demonstrated that some topics have the same terms or very close terms to define them. For this reason, we tried to modify the number of topics and find the best model to define the findings being analyzed.

Figure 2 shows the same kind of visualization as in Figure 1, but with 10 topics. The number of overlaying topics decreased and the distance between them increased. These facts show that the model improved. However, even though this model is better than the previous one, some topics are still overlaying such as 1 and 3, 6 and 8, 9 and 10, and 2 and 5. In an good model, there should have no overlaying and topics should not be too close to each other. This means that the topics are more likely to represent different concepts/domains.

Figure 3 displays the best result achieved, with 5 topics. Here we can see that there is no overlaying. On the one hand, the models with more topics we tested before (20, 15, 12, 10, 7, and 6) have some overlapping. On the other hand, the models with less topics, such as 4, tends to put together different concepts/domains in the same topic. In conclusion 5 seems to be the best model for identifying the topics in the IT audit findings.

The experiment previously described applied LDA to the detailed text column. We also compared it with the same model using the resumed text column. In order to validate the result, the top 15 terms in each model (resumed and detailed text) were compared. Despite presenting similar terms, the detailed text

[4] https://github.com/cpsievert/LDAvis



Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 1. LDA Visualization with 20 Topics

presented more consistent/similar words per topic than using the resumed text, according to the SMEs.

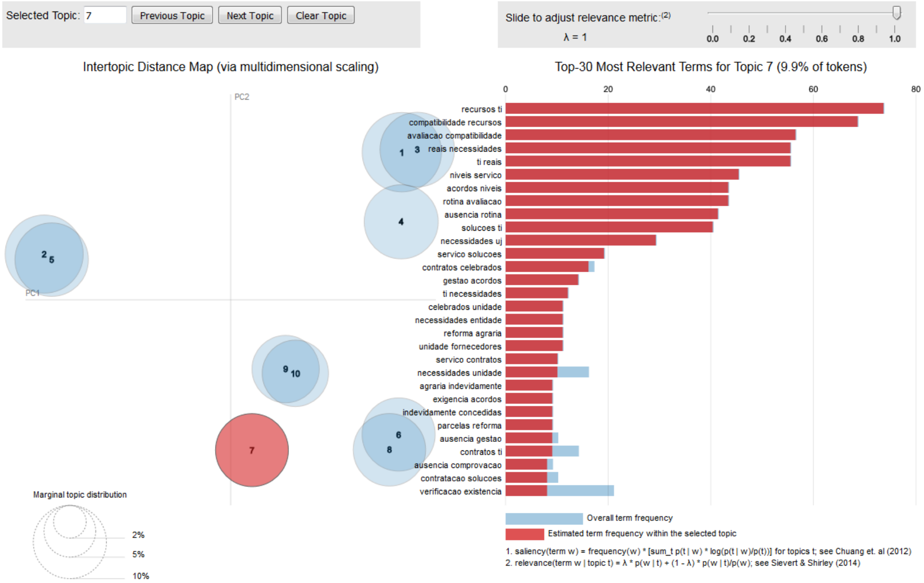
The topics found in the best model can be resumed as:

- Topic 1: Goods and services, contracts.
- Topic 2: Human resources, career
- Topic 3: Strategic planning, PDTI
- Topic 4: IT security, public security
- Topic 5: Software developing, function point

After selecting the best model, we can see one specific document and the probability of the topics inside it. LDA separates the clusters checking the frequency of words or terms in a document and comparing this frequency with the words or terms that is defined in each topic. In general, the documents have words that appear in different topics. Therefore, according to the frequency of these terms, it calculates the proportion that one document would have of belonging to each of the topics defined by the model. The LDA will choose the topic which presents the highest probability.

Figure 4 displays an example of probability distribution of topics inside a document. This document belongs to topic 3 because the terms of this topic are more frequent. Note, however, that the probabilities do not have to sum to 1.

Figure 5 illustrates an example based on part of the detailed text column after pre-processing. The terms present in each topic were highlighted with the



Selected Topic: 7

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. LDA Visualization with 10 Topics

colors of the topics they belong to. Looking at the colored text it is easy to see that this text belongs to topic 3 because the most predominant color is green (the color assigned to topic 3).

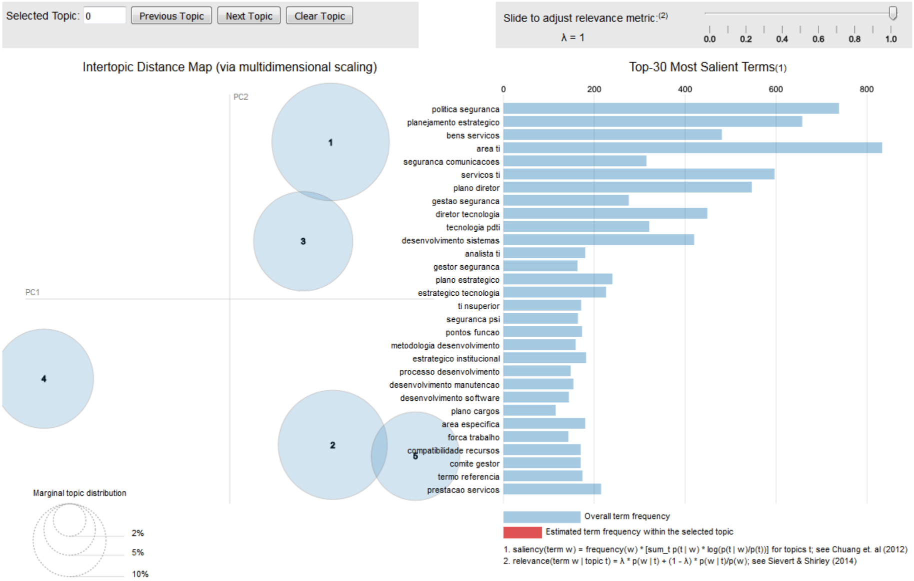
3.2 Supervised Learning

In order to complement and validate the results of the topic modeling experiment described in Section 3.1, we built a supervised learning model, using a database previously classified by experts. This dataset was created from a sample of 335 records from the main database. This number represents 13.4% of the registers and was calculated in order to allow a 95% confidence level and a 5% margin sampling error.

The field used to support the classification was the "detailed text". We emphasize that the categories were created without prior knowledge (by the experts) of the topics found in the previous approach.

The experts found eight categories of findings in the sample. Table 1 shows the distribution of these findings by category found.

The category "IT Management" involves noncompliance with the Brazilian standards that determine a minimum configuration of IT departments in government units. "IT Contracts" refers to problems found in IT procurement and hiring processes. "IT Human Resources" refers to failures in personnel policy es-



Selected Topic: 0

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 3. LDA Visualization with 5 Topics

tablished for the IT areas. “Information Security Policy” relates to the absence of clear rules of information security. “Software Development and Maintenance” involves problems like delays in the implementation of systems or the lack of use of the implemented system. “Internal Control” covers the problems in the process for assuring achievement of the IT department objectives. “Outsourcing” refers to specific contracts involving hiring IT personnel. “Not specified” is a category created to the registers where there is not enough terms to explain its meaning.

As shown in Table 1, there are too few registers in categories “Internal Control” and “Outsourcing”. An alternative to get better results in machine learning processes is to join two or more categories (an example can be found in ALEJO [1]). At first, in this specific case we preferred to keep them in separate categories, as it is not so clear their relationship with other categories.

After creating the labeled database we provided a transformation in the data by converting the field “detailed text” into a term frequency inverse document frequency (TF-IDF) vector. This vector represents a numerical statistic that increases proportionally to the number of times a word appears in a specific document, but is offset by the frequency of the word in the other documents [17].

We built the Machine Learning model using a Random Forest algorithm. This algorithm implements a Random Decision Forest, that is an ensemble learning

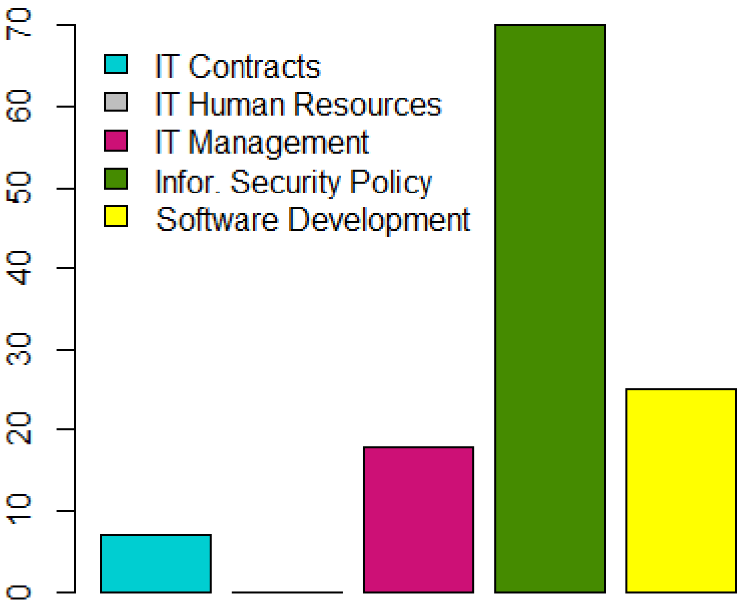
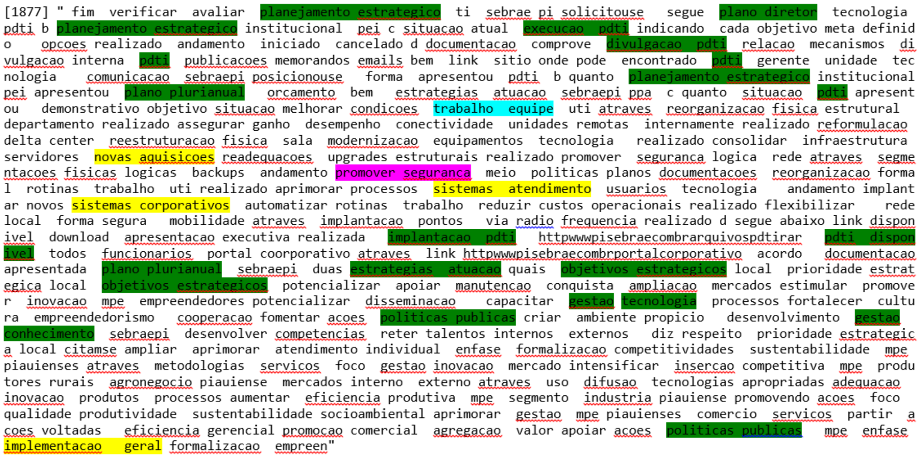


Fig. 4. Distribution of Topics in a Document



[1877] " fim verificar avaliar planejamento estrategico ti sebrae pi solicitouse segue plano diretor tecnologia pdti b planejamento estrategico institucional pei c situacao atual execucao pdti indicando cada objetivo meta definido opcoes realizado andamento iniciado cancelado d documentacao comprove divulgacao pdti relacao mecanismos divulgacao interna pdti publicacoes memorandos emails bem link sitio onde pode encontrado pdti gerente unidade tecnologia comunicacao sebraepi posicionouse forma apresentou pdti b quanto planejamento estrategico institucional pei apresentou plano plurianual orcamento bem estrategias atuacao sebraepi ppa c quanto situacao pdti apresentou demonstrativo objetivo situacao melhorar condicoes trabalho equipe uti atraves reorganizacao fisica estrutural departamento realizado assegurar ganho desempenho conectividade unidades remotas internamente realizado reformulacao delta center reestruturacao fisica sala modernizacao equipamentos tecnologia realizado consolidar infraestrutura servidores novas aquisicoes readequacoes upgrades estruturais realizado promover seguranca logica rede atraves segmentacoes fisicas logicas backups andamento promover seguranca meio politicas planos documentacoes reorganizacao formal rotinas trabalho uti realizado aprimorar processos sistemas atendimento usuarios tecnologia andamento implantar novos sistemas corporativos automatizar rotinas trabalho reduzir custos operacionais realizado flexibilizar rede local forma segura mobilidade atraves implantacao pontos via radio frequencia realizado d segue abaixo link disponivel download apresentacao executiva realizada implantacao pdti httpwwwpisebraecombrarquivospdtirar pdti disponivel todos funcionarios portal coorporativo atraves link httpwwwpisebraecombrportalcorporativo acordo documentacao apresentada plano plurianual sebraepi duas estrategias atuacao quais objetivos estrategicos local prioridade estrategica local objetivos estrategicos potencializar apoiar manutencao conquista ampliacao mercados estimular promover inovacao mpe empreendedores potencializar disseminacao capacitar gestao tecnologia processos fortalecer cultura empreendedorismo cooperacao fomentar acoes politicas publicas criar ambiente propicio desenvolvimento gestao conhecimento sebraepi desenvolver competencias reter talentos internos externos diz respeito prioridade estrategica local citamse ampliar aprimorar atendimento individual enfase formalizacao competitividades sustentabilidade mpe piauienses atraves metodologias servicos foco gestao inovacao mercado intensificar insercao competitiva mpe produtores rurais agronegocio piauiense mercados interno externo atraves uso difusao tecnologias apropriadas adequacao inovacao produtos processos aumentar eficiencia produtiva mpe segmento industria piauiense promovendo acoes foco qualidade produtividade sustentabilidade socioambiental aprimorar gestao mpe piauienses comercio servicos partir acoes voltadas eficiencia gerencial promocao comercial agregacao valor apoiar acoes politicas publicas mpe enfase implementacao geral formalizacao empreen"

Fig. 5. Document with Terms Highlighted by Topic

Table 1. Quantity of Findings per Category

Finding Category	Quantity
IT Management	104
IT Contracts	72
IT Human Resources	42
Information Security Policy	41
Software Development and Maintenance	31
Internal Control	10
Outsourcing	9
Not specified	26
Total	335

technique for classification [11,7], that works by constructing a set of decision trees that depend on the values of a random vector sampled independently and with the same distribution of all trees in the forest [5].

In order to start the learning process we split the database into two subsets. The first (70% of the registers) one was allocated to estimate and calibrate the model, using a 10-fold cross-validation process. This technique is commonly used to calibrate predictive models and involves partitioning the training data into ten subsets, performing the analysis on one subset and validating the analysis on the other subset. This process is repeated in multiple rounds of cross-validation using different partitions, in order to reduce variability. The second subset (30% of the registers) was used as a test dataset.

The cross-validation process shows the set of trees in the Random Forest model that best fit the training data. After this step we got an accuracy [15] of 74.33%, which means that on average 74.33% of the records were classified correctly in the iterations. Another important indicator to measure the model efficiency is the Kappa index [6], which compares the overall accuracy with that obtained in each class, and is helpful to differentiate the prediction results from those that would be obtained randomly. We obtained a 0.67 score, considered "substantial" in accordance with the existing guidelines for the interpretation, as shows in Table 2 [12].

Table 2. Values of Kappa x Levels of agreement

Value of Kappa	Level of Agreement
less than 0.01	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost Perfect agreement

Applying the calibrated model in the test subset we got an accuracy of 73.27% and a Kappa index of 0.66. It is common in such classification models to loose some prediction capability when comparing training and testing. Nevertheless the results were still very impressive and not statistically different than the results obtained during cross-validation.

A more accurate analysis of the errors shows us that most incorrect classifications occurred in "Internal Control" and "Outsourcing" categories. This result was expected, since these are the classes with few occurrences.

4 Conclusions

In this paper we investigated the IT audits of the Brazilian Office of Comptroller General. We applied a supervised (text classification) and unsupervised (LDA) approach. Both of them achieved very similar results.

Comparing both approaches, we can see a reasonable congruence between modeling topics (unsupervised) and the model learned from the manual classification by the specialist. In fact, the topics discovered in the first approach have high correlation with the manually identified classes used in the learning model. Figure 6 shows the correlation between the topics found (represented here by the most frequent words) and the classes defined in the supervised learning model.

As we can see, the five topics found in the best model using LDA were all represented in the classification model. The only classes that we could not related to the learned topics were the least frequent ("Outsourcing" and "Internal

Controls"), probably because they represent only 5% of the ratings. The reason for this is that the topic modeling presented better results using 5 topics, which meant that only the most representative categories of supervised model met correspondence.

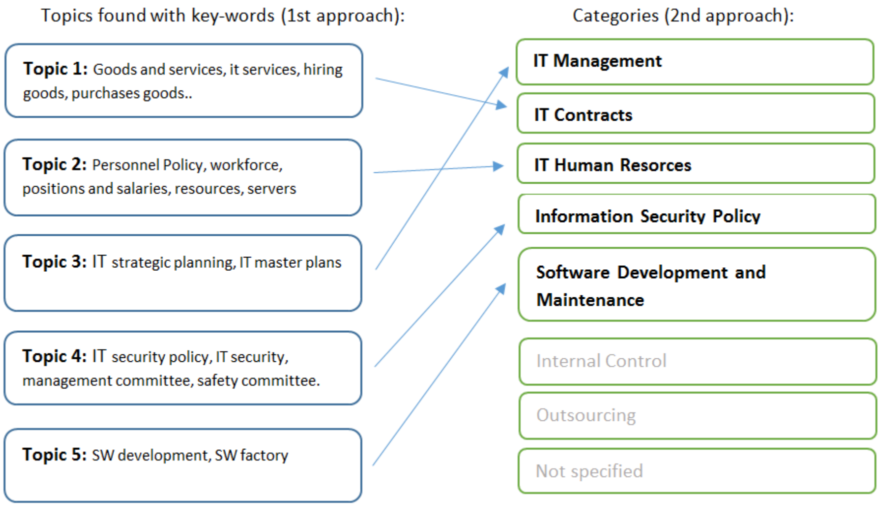

Topics found with key-words (1st approach):
Categories (2nd approach):
Topic 1: Goods and services, it services, hiring goods, purchases goods..
Topic 2: Personnel Policy, workforce, positions and salaries, resources, servers
Topic 3: IT strategic planning, IT master plans
Topic 4: IT security policy, IT security, management committee, safety committee.
Topic 5: SW development, SW factory
IT Management
IT Contracts
IT Human Resorces
Information Security Policy
Software Development and Maintenance
Internal Control
Outsourcing
Not specified

Fig. 6. Topic x Categories

This is the first study that analyzes CGU's audit reports using a well-known knowledge discovery methods. We believe that this study may be expanded to cover other types of audits (since here we covered only audits of IT resources). Another possible further development is to understand (*e.g.*, through a supervised learning model) the main problems that contribute to a finding being considered or classified as severe.

The great benefit of this work is the ability to learn from experience by using data mining techniques. A better understanding of the problems can be used for planning future audits or even for reshaping the audit proceedings used, in order to increase the effectiveness of the CGU's performance.

Acknowledgments The authors would like to thank the Federal Internal Control Secretariat (SFC, in Portuguese), specially the Coordination of Control and Core Audit of Information Technology (GSNTI, in Portuguese) for the partnership during this work. Finally the authour thank the Director of the Department of Research and Strategy Information, Gilson Libório, and CGU for their support and for allowing the publication of this work.

References

1. R. Alejo, R. Valdovinos, V. Garca, and J. Pacheco-Sanchez. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, 34(4):380–388, Mar. 2013.
2. D. Andrzejewski and D. Buttler. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 600–608. ACM, 2011.
3. M. W. Berry and M. Castellanos. *Survey of text mining II.* Springer, 2008.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
5. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
6. J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.
7. H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 169–178. IEEE, 2008.
8. Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. 2002.
9. D. Ganguly and G. J. Jones. Partially Labeled Supervised Topic Models for RetrievingSimilar Questions in CQA Forums. pages 161–170. ACM Press, 2015.
10. D. Jurasfsky and J. H. Martin. *Speech and Language Processing.* Stuart Russell and Peter Norvig, 1998.
11. V. Korde. Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85–99, Mar. 2012.
12. J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, Mar. 1977.
13. P. Maia, R. N. Carvalho, M. Ladeira, H. Rocha, and G. Mendes. Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency.
14. C. A. Martins, M. C. Monard, and E. T. Matsubara. Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proceedings of The Third IASTED International Conference on Artificial Intelligence and Applications (AIA 2003), Benalmdena, Espanha.(to be published)*, volume 38, 2003.
15. Michael Gordon and Manfred Kochen. Recall-precision trade-off: A derivation, 1988.
16. S. Pyo, E. Kim, and M. kim. Lda-based unified topic modeling for similar tv user grouping and tv program recommendation. *IEEE Transactions on Cybernetics*, 45(8):1476–1490, Aug 2015.
17. J. Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.
18. L. Sales. Risk prevention of public procurement in the brazilian government using credit scoring. Obegef working papers, OBEGEF - Observatrio de Economia e Gesto de Fraude, 2013.
19. C.-M. Tan, Y.-F. Wang, and C.-D. Lee. The use of bigrams to enhance text categorization. *Information processing & management*, 38(4):529–546, 2002.
20. X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.