EVALUATION OF TOPIC MODELS IN CONSUMER RESEARCH.

Abstract: The need for data compression and analysis tools has been increasingly crucial as a lot of data is being generated every minute with the continuous growth of several sources of information. Topic modeling has become a useful tool for the text analysis, especially the Latent Dirichlet Allocation (LDA) model. This study aims to evaluate the classification of open responses in a customer satisfaction survey using the LDA model. This study compared the classification of customers' responses using the LDA model with that of three researchers. A responde classification accuracy higher than 80% was identified in this study. The LDA technique was able to separate the attributes and emerging words, which were highlighted as problems to be solved according to supermarkets' managers.

Keywords: Latent Dirichlet Allocation, Text analysis, Accuracy, Topics extraction.

1 INTRODUCTION

It took the radio 38 years to reach an audience of 50 million users, television took more than a decade (13 years), and it took the Internet 3 years to get 50 million users. The data become more impressive if we consider social networks. Facebook reached 50 million users after one year of its release in 2004 while Twitter reached the same figure in nine months. These numbers show that the media is just one example of the big data production (Chui, Manyika & Bughin, 2012).

Companies in almost all sectors are focused on data exploration mainly as a source to have competitive advantage with the large amount of data that is available. However, Davenport (2014) highlights that the point is not to be dazzled by the volume of data, but rather to analyze it, to convert it into insights, innovations and business value.

Regarding the text data, Srivastava, Salakhutdinov and Hinton (2013) emphasize that these data are originally available in an unstructured way. It is necessary to use algorithms that are able to discover patterns and trends through topic modeling in order to have a useful representation of documents and high-quality information.

The main goal of topic modeling is to discover the main subjects that permeate a collection of documents; it means to discover patterns in the use of words and how to connect documents that share similar patterns. Researchers have published many studies about topic modeling applied in several fields such as software engineering, political science, medical science, linguistics, business and management. Business and management are fields in which the main focus is on information about the level of customers' satisfaction (Balducci & Marinovacci, 2018).

There are several applications of topic models in customer satisfaction surveys: for example, Gao, Yu and Liang (2016) used topic modeling to analyze 17,747 comments from public transport customers in the United States. According to them, the most frequently identified customer satisfaction attributes are: waiting time, cleanliness, accessibility, comfort. Lucini, Tonetto, Fogliatto and Anzanello (2020) analyzed 55775 Online Customer Reviews (ORCs) from passengers from different countries and airlines and identified 27 dimensions of customer satisfaction. Liao, Chen, Ku, Narula and Duncan (2020) focused on insurance companies in the USA and they used topic modeling to analyze information from customer calls. They classify and process the information more efficiently by using approximately 10,000 customer calls to develop the study. Sutherland, Sim, Lee, Byun and Kiatkawsin (2020) used 104,161 online reviews from Korean accommodation customers to identify which topics are considered important by guests. The researchers identified 14 topics such as accessibility, hospitality, room size, among other things. The main or most popular of the probabilistic topic models, the Latent Dirichlet Allocation (LDA), was used by the authors in all the studies presented here.

Topic models are commonly used to classify and analyze a large volume of text documents. Several metrics have been proposed to measure the coherence of the topics. Statistical metrics, such as perplexity, are used to adjust models (Mimno, Wallach, Talley, Leenders & McCallum 2011; Newman, Lau, Grieser & Baldwin, 2010).

Chang, Gerrish, Wang, Boyd-Graber and Blei (2009) presented the first evaluation of topic models by humans, but the study focused on the interpretation of topics and not on the assertiveness of the classification. They also emphasized that further research should be done in the development of topic models that focus on evaluations compared to the real world, rather than optimizing probability-based measures. Bache, Newman and Smyth (2013) highlighted that the first words that characterize a topic cannot be considered a coherent summary of the large number of words. Newman, Asuncion, Smyth and Welling (2009) emphasized that some topics learned by a model are not interpretable and useful for human use, although statistically coherent.

This study aims to evaluate the classification of open responses in a customer satisfaction survey by using the LDA model. This is how this study is organized: chapter 2 presents the review of literature on the LDA Model; chapter 3 presents the methodology used to achieve the objectives of this study; chapter 4 presents the results, and chapter 5 presents the conclusions of this research.

2 LATENT DIRICHLET ALLOCATION

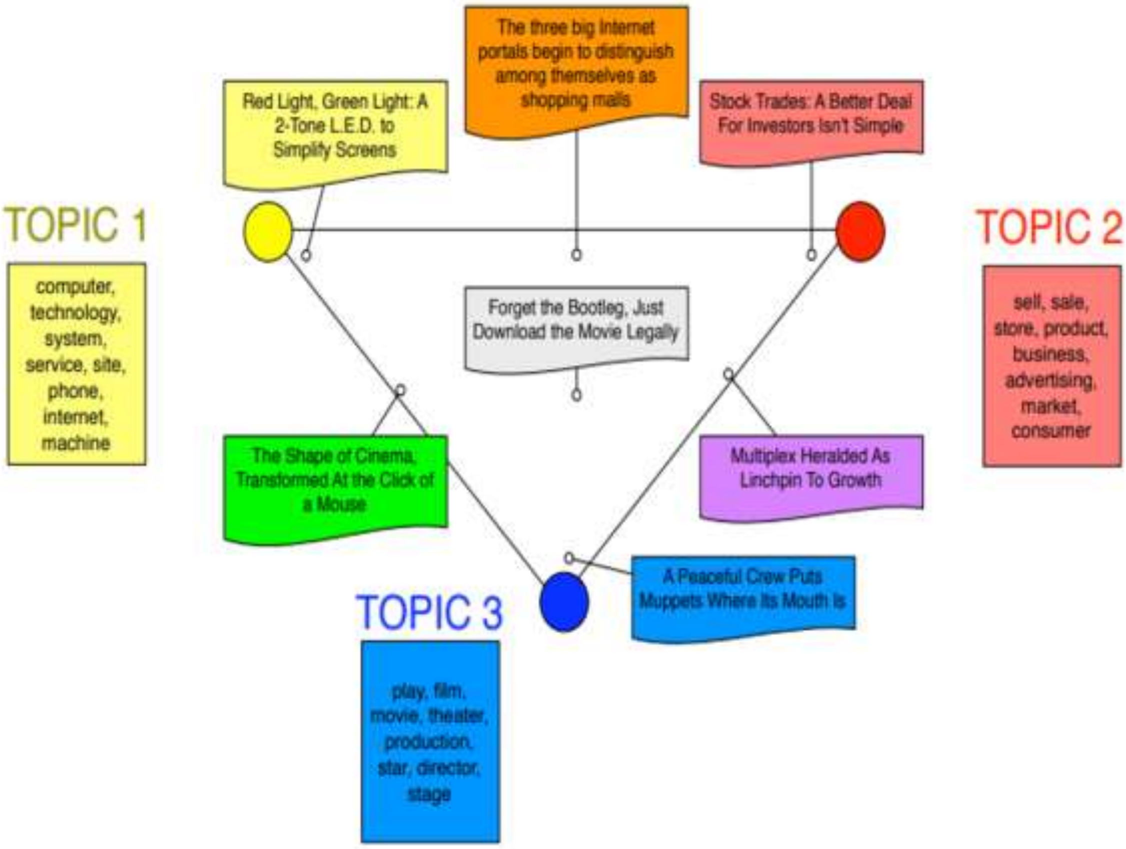
A topic model is based on corpus of documents. It discovers the topics that permeate the corpus and assigns documents to those topics. Thus, a topic model can be thought as a box with the following output: assigning words to topics and assigning topics to documents (Han, Pei & Kamber, 2011).

The first output is the distribution about words, as shown in Figure 1, which shows three uncovered topics (Topic 1, Topic 2 and Topic 3). The topics are usually presented as word lists, or a bunch of words. This is usually enough to obtain an approximate understanding of the topic (Chang et al., 2009).

Documents are assigned to topics in the second output of a topic model. This step is represented by the simplex, in which each of the ten documents is associated with the three topics, showing the position of each document in the topic space in Figure 1. There is a wide variety of methods to find these topics and perform the assignment of topics to documents, one of the most quoted in the literature is the LDA (Chang et al., 2009).

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. Observed variables are the terms of each document and unobserved variables are topic distributions. The parameters of the topic distributions, also known as hyperparameters, are given in the model. The distribution that is used is the Dirichlet. The Dirichlet sampling result is used to allocate words from different topics that will fill the documents in the generative process. Thus, the meaning of Latent Dirichlet Allocation, which expresses the intention of the model to allocate latent topics that are distributed in accordance with the Dirichlet distribution, can be oberved (Blei, NG & Jorndan, 2003).

Figure 1. Topics and assignment of documents to topics.



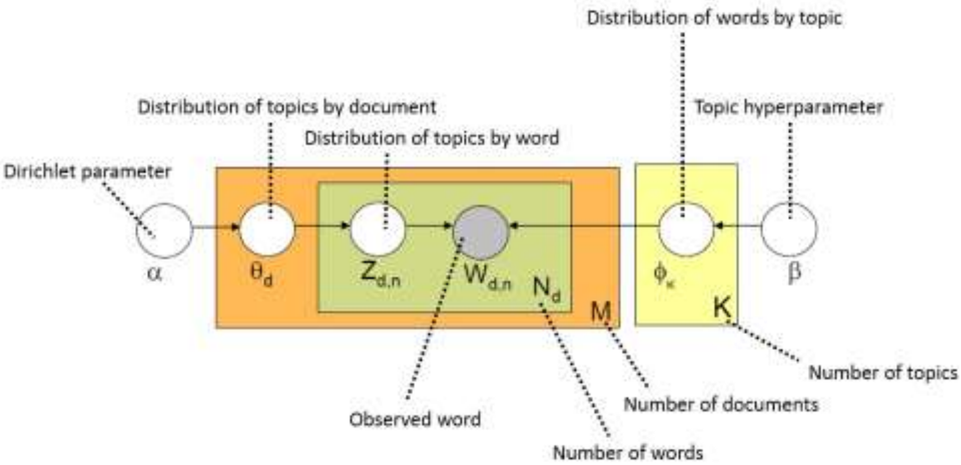
The three big Internet portals begin to distinguish among themselves as shopping malls
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Stock Trades: A Better Deal For Investors Isn't Simple
TOPIC 1
computer, technology, system, service, site, phone, internet, machine
TOPIC 2
sell, sale, store, product, business, advertising, market, consumer
Forget the Bootleg, Just Download the Movie Legally
The Shape of Cinema, Transformed At the Click of a Mouse
Multiplex Heralded As Linchpin To Growth
A Peaceful Crew Puts Muppets Where Its Mouth Is
TOPIC 3
play, film, movie, theater, production, star, director, stage

Source: Adapted from Chang et al. (2009).

The intuitive idea beyond the LDA has been showed by some authors (Blei et al., 2003). It is assumed that a number of topics, which are distributions about words, are observed in the entire collection of documents. Each document is assumed as being generated as follows: first choose a distribution over the topics; then, for each word, choose a topic assignment and choose the word from the corresponding topic. The author exemplified the idea by using a scientific paper about the use of data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense). By analysing the paper, words such as computer, prediction, life, organism, genes, and sequenced appear. If there were time to highlight each word of the paper, it could be verified that it blends genetics, data analysis and evolutionary biology in different proportions. LDA is a statistical model of document collections that attempts to capture this intuition (Blei, 2012).

The generative process of LDA is imaginary and the opposite of what is proposed in a computational task of data mining. In the LDA process, it is assumed that the topics are defined before any data can be generated. In this study, the topics are defined as probabilistic distributions over a fixed set of words. The documents are bag of words that arise from the probabilistic choice of words that belong to a distribution of topics. All the generative process may be represented by a Bayesian network. This network is shown in Figure 2.

Figure 2. LDA graphical model



Source: Adapted from Blei et al. (2003).

As previously presented, the Bayesian model of LDA is hierarchical and it has three levels: the first level represents the distribution of topics in the entire collection of documents. The second level brings the distribution of topics for each document. The third level consists of repeating the internal distribution of topics to the words in a document. This last level enables the representation of a document as a blend of topics.

In order to represent the distributions, two variables are used: ϕ is an n-dimensional variable in which n is the number of words over the vocabulary. The θ is a K-dimensional variable in which K is the number of topics. These two variables are generated by the Dirichlet distribution with their respective β and α hyperparameters (Blei, 2012).

In LDA, the same document may be related to several topics with different proportions of relevance since each document has its own distribution of θ_d topics. This can be seen in the generative model by choosing the topic assigned to the variable z_{dn}, in which there will occasionally be a chance to choose different topics according to the θ_d distribution (Blei, 2012).

Considering the observed and non-observed variables, this study aims to find out the assignment of topics to documents, the distribution of documents per topic, and topics per vocabulary. It means that the computational problem of LDA is to infere $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|\, w_{1:D})$, in which w represents all the observed words in the collection of documents (Blei et al., 2003). According to Bayes, it may be possible to formulate the probability as the calculation of the posterior of the LDA. Thus, there is equation 1: the numerator is the joint distribution of all non observed and observed variables and the denominator is the marginal probability of the observations.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|\, w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \qquad (1)$$

Thus, the main computational problem may be solved by infering the probability the posterior from all the model, which was described in Equation 1. This may be thought as the reverse of the generative process. In theory, this inference may be done by the sum of the joint distribution of all possible values assigned as non-observed variables (all words in the collection). The number of possible attributions is exponentially large, what makes this intractable to compute. (Blei, 2012). There are several methods to approximate the posterior distribution and the most used method in the literature for inference of the LDA model is the Gibbs Sampling (Griffiths & Steyvers, 2004).

Gibbs sampling algorithm for topic modeling is the most popular mainly due to its easy implementation and function. This is the case of the Monte Carlo Markov chain simulation. Monte Carlo methods in Markov chain can emulate probability distributions with high dimensionality through the stationary behavior of the Markov chain (Geman & Geman, 1984).

3 METHOD

To achieve the purpose of this study, data from a survey conducted with Supermarket customers (described in detail in Chapter 3.1) and the methodological procedures and records detailed in chapter 3.2 were used.

3.1 DATA SET

The data used in this study are part of a research carried out with two supermarket branches in a city located in the south of Santa Catarina, which are referred here as branch A and branch B. The survey was carried out in residences located in the neighborhoods where the branches were located. It was possible to dimension a sample of 400 questionnaires, with an estimation error of 4.88% and with a 95% accuracy level in each neighborhood based on the data collected from Censo Demográfico 2010 (IBGE, 2018).

Data collection was carried out in September 2018 through a structured questionnaire containing pre-defined closed and open questions. Only open questions will be considered in this study. (Table 1). For the selection of the respondents, it was considered the person who mostly shopped in the supermarket.

Table 1. Questions.

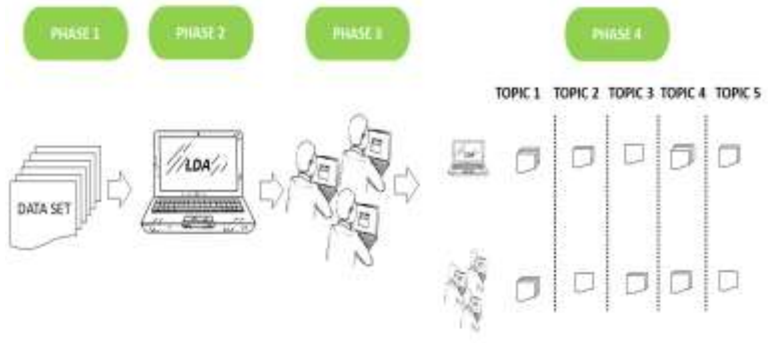
Q1	What things does the supermarket do and it should keep doing?
Q2	What things does the supermarket do and should stop doing?
Q3	What things should the supermarket do that it doesn’t?

Source: Authors.

3.2 METHODOLOGICAL PROCEDURES

The methodological procedures used to achieve the objectives of this study were divided into three phases and they are presented in Figure 3 and explained later.

Figure 3. Phases



Source: Authors.

Phase 1 – The set of raw data was text pre-processed; special characters and stopwords were removed from the questionnaires (400 from each branch).

Phase 2 – The file with the data base was processed by using the LDA model. It was decided to use K = 5 (number of topics or categories) after a previous analysis of the results. The Griffrs 2004 metrics was used to finish the K parameter, and the Gibbs Sampling algorithm with 1000 repetitions and burn-in of 500 repetitions was used.

Phase 3 – The file with the answers and a list of five words representing each category was sent to three researchers who are Business graduate professionals. They were asked to classify each answer in just one category and to write down the time needed to perform this task.

Phase 4 – The results were compared to the LDA processing once the files were sent back by the researchers. The results from the analysis of the LDA were compared to those classified by the researchers in order to calculate the LDA accuracy. The accuracy corresponds to the proportion of answers that were equally classified (Researcher and LDA) considering all the answers.

The text analysis and topic modeling package from RStudio Software was used to perform the pre-processing of text and the LDA. The answers were sent to the researchers in Excel for them to have a clear view.

4 RESULTS AND DISCUSSION

The phases used in text pre-processing were very similar to those used in previous studies (such as Xiyue). When proceeding to the practical application and interpretation of the results, there was an average of 116 words in each question. The number of words is very limited when compared to studies by Lucini et al. (2020) and Sutherland et al. (2020), with thousands of words available for analysis.

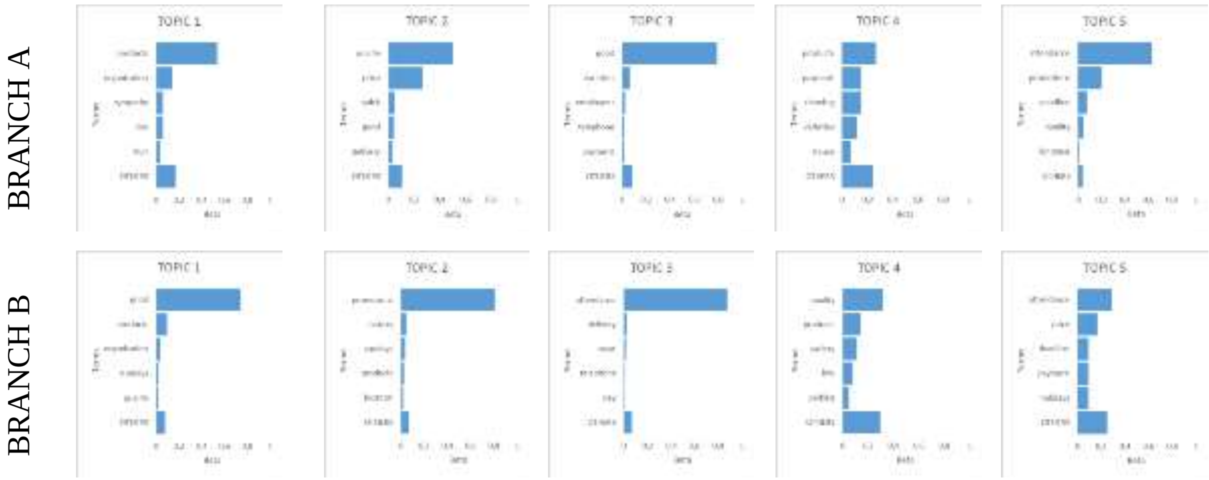
There is a predominance of the words such as “service” and “good” at both branches. Words such as “products”, “quality” and “offer” emerged at Branch A in the analysis of words in Q1. The question 2 represent the negative aspects pointed out by costumers. Words such as “line”, “meat”, “vacuum” and “price” were predominant at Branch A. At Branch B, words such as “bad”, “service”, “delivery” and “card” emerged. The word both Branches had in common was “line”.

Answers to question 3 (Q3), there is a predominance of the words “cashier”, “movement”, “per” and “days” at Branch A. On the other hand, Branch B had words such as “parking”, “offers”, “fruits” and “improve”.

The responses found at Branch B are more evenly distributed than at Branch A's responses, which explains a more sparse word cloud and indicates that Branch B has more problems to be solved.

The LDA model was used to identify the five emerging topics and classify the respondents’ answers in the phase of topics extraction. The words that are more likely to be in each of the five topics (the same that were sent to the researchers) are shown in figure 3. These words refer to Q1 at branch A, and they refer to service, organization, quality, price, offers and products. The extracted topics that refer to Q1 at branch B refer to service, quality, price, offers and products.

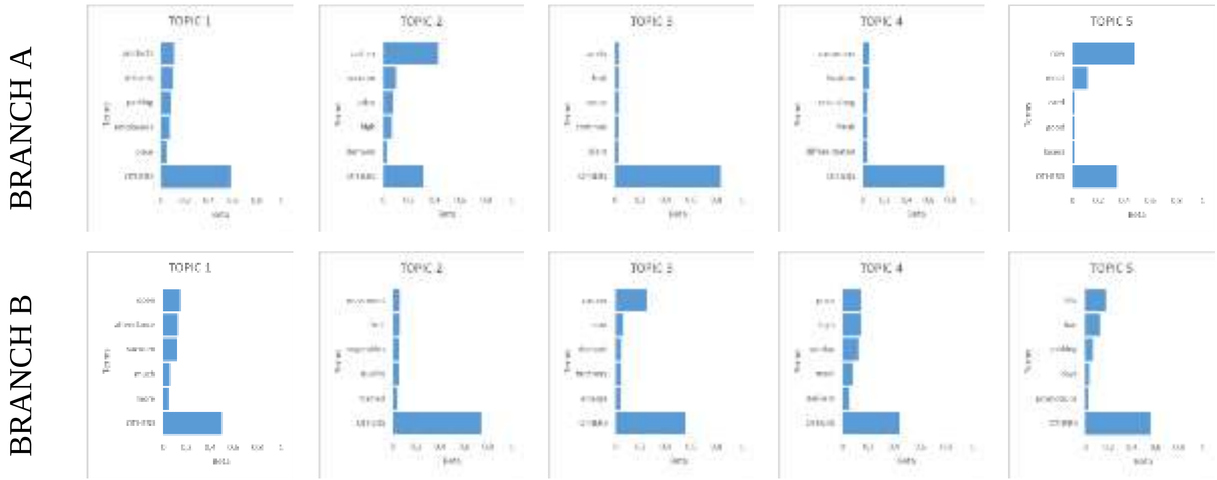
Figure 4. The most probable words from the topics – Q1 (branch A and branch B)



Source: Authors

Topics that refer to Question 2 from Branch are shown in figure 5. These topics refer to products, cashier, meat and line. The topics that refer to Q2 at branch B refer to cashier, price, line and parking.

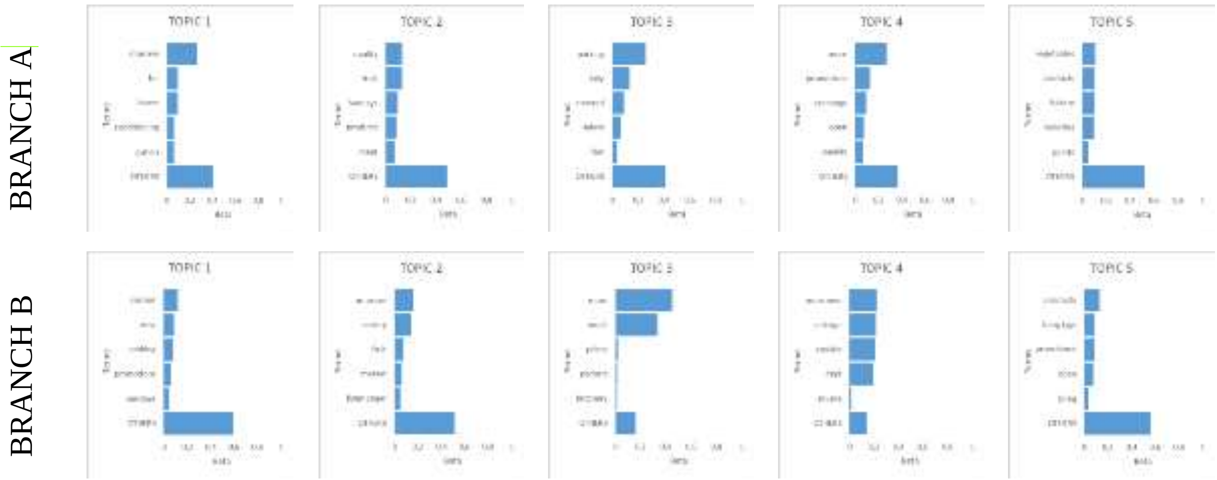
Figure 5. The most probable words from the topics – Q2 (branch A and branch B).



Source: Authors

The topics from Q3 of branch A are shown in figure 6. The topics extracted refer to promotions, boxes, parking and quality. The topics from Q3 of branch B refer to boxes, fruits, parking, packers and service.

Figura 6. The most probable words from the topics – Q3(branch A and branch B).



Source: Authors

The results of the classifications of the LDA and each researcher at both branches and question (accuracy result) were compared and are shown in table 2. Researcher 3 presented the best average accuracy, 86.33%, it means that for every 100 responses classified by this researcher, 86 received the same classification as the LDA model. Researcher 2 had the worst average accuracy, 83.92%.

Table 2. Result of accuracy

Researcher	Branch and Question	Accuracy (%)	Accuracy (%) Average	Accuracy (%) Question Average	Accuracy (%) Branch Average
Researcher 1	A_Q1	80,25	86,21	Q1	Branch A
	A_Q2	88,50		78,75	86,67
	A_Q3	91,25		Q2	
	B_Q1	77,25		90,38	Branch B
	B_Q2	92,25		Q3	85,75
	B_Q3	87,75		89,50	
Researcher 2	A_Q1	73,25	83,92	Q1	Branch A
	A_Q2	85,00		74,00	82,58
	A_Q3	89,50		Q2	
	B_Q1	74,75		88,13	Branch B
	B_Q2	91,25		Q3	85,25
	B_Q3	89,75		89,63	
Researcher 3	A_Q1	92,00	86,33	Q1	Branch A
	A_Q2	83,75		83,38	88,25
	A_Q3	89,00		Q2	
	B_Q1	74,75		84,38	Branch B
	B_Q2	85,00		Q3	84,42
	B_Q3	93,50		91,25	

Source: Authors

The accuracy of Question 1 by researchers 1 and 2 was the lowest (78.75% and 74.00%, respectively). It is possible to identify words that are difficult to categorize through an analysis with the list of words sent to the researchers because they refer to very close categories. In relation to the average per branch, there is an accuracy of around 85%.

It is important to understand whether the difference in accuracy found between the result of the classification in the LDA and the researchers is acceptable since there is a lot of subjectivity of the researcher when it comes to the classification of responses. For this reason, each classification was evaluated individually and the amount of times that the same answer was equally classified by all researchers was verified. An average result of 89.96% was obtained (Table 3).

Table 3. Result of accuracy researchers

Branch and Question	Same answers among Researchers (%)	Average (%)
A_Q1	72,25	89,96
A_Q2	91,50	
A_Q3	97,75	
B_Q1	91,50	
B_Q2	92,50	
B_Q3	94,25	

Source: Authors

Qiang, Qian, Li, Yuan and Wu (2020) carried out a study with some models of text classification in different databases. The best performance of the LDA model was in the base for twets, with a little more than 80% accuracy. However, the accuracy in the biomedicine base did not reach 50%. Their study focused on STTM (short text topic modeling) and the data to perform the comparison was performed with machine learning.

Towne, Rosé and Herbsleb (2016) conducted a study to evaluate the similarity between LDA and Human Perception. They noted that for the most part, under deliberately chosen favorable conditions for observing concordance, judgments agree at approximately 75% and other conditions at approximately 66%.

There was a conversation with the managers of each branch after the results were obtained in order to understand what happens at the branches. At branch A, the manager reported that one of the biggest problems is the checkout line, as evidenced by the results. Branch B manager points out that the unit is new in the neighborhood (less than 1 year when the survey was carried out) and that there are some problems resulting from adapting to the neighborhood, such as the parking lot (which was being built) and the people who work at the branch has little experience.

Time was a very significant factor when we compare researchers and LDA. The processing time (classification only) to perform the classification via LDA was approximately 1 minute. The processing was performed on a 2.4 GHz Intel I-7 Core (TM) processor with 8 GB of RAM. The time to perform the classification task by researchers 1, 2 and 3 was 18h32min, 14h32min and 15h21min respectively.

5 CONCLUSIONS

As the goal of this study was to evaluate the use of the LDA model for processing and analyzing non structured open questions in a small data set, it is evident the amount of time saved for processing this type of data. The main limitation of our approach is the reduced number of researchers to carry out the classification, and use of only one database.

The tools (software) that are most commonly used for treatment and analysis of open questions make the work of the researcher hard as it is necessary to read and interpret each answer, besides resulting in the researcher's subjectivity, which could be demonstrated when comparing the classifications among Researchers.

It was possible to classify the 800 questionnaires in approximately 1 minute by using the LDA model. This technique was able to separate the attributes, the most emerging words and the words that were identified by supermarket managers as potential problems to be solved. For further studies, continuous monitoring of customer satisfaction is recommended not only for the supermarket sector. Despite the significant sample size of this research, the results obtained can be considered standard and generalized only for the researched branches.

REFERENCES

Bache, K., Newman, D., & Smyth, P. (2013, August). Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 23-31).

Balducci, B., & Marinova, D. (2018). *Unstructured data in marketing. Journal of the Academy of Marketing Science*, 46(4), 557-590.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).

Chui, M., Manyika, J., & Bughin, J. (2012). *The social economy: Unlocking value and productivity through social technologies*. McKinsey Global Institute.

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities.* Harvard Business Review Press.

Gao, L., Yu, Y., & Liang, W. (2016). Public transit customer satisfaction dimensions discovery from online reviews. *Urban Rail Transit*, 2(3-4), 146-152.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228-5235.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

IBGE (Instituto Brasileiro de Geografia e Estatística). Censo Demográfico, 2010. Available: http://www.sidra.ibge.gov.br. Access: Jun 2018.

Liao, X., Chen, G., Ku, B., Narula, R., & Duncan, J. (2020). Text Mining Methods Applied to Insurance Company Customer Calls: A Case Study. *North American Actuarial Journal*, 24(1), 153-163.

Lucini, F. R., Tonetto, L. M., Fogliatto, F. S., & Anzanello, M. J. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *Journal of Air Transport Management*, 83, 101760.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272).

Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(8).

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100-108).

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering.*
Srivastava, N., Salakhutdinov, R. R., & Hinton, G. E. (2013). Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865.*

Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). *Topic modeling of online accommodation reviews via latent dirichlet allocation. Sustainability,* 12(5), 1821.

Towne, W. B., Rosé, C. P., & Herbsleb, J. D. (2016). Measuring similarity similarly: Lda and human perception. *ACM Transactions on Intelligent Systems and Technology (TIST),* 8(1), 1-28.