Topic Modeling for Auditing Purposes in the Banking Sector

Alessandro Giaconia[1,*], Valeria Chiariello[2], Sara Giannuzzi[2] and Marco Passarotti[1]

[1]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italy
[2]CREDEM, Via Emilia San Pietro 4, 42121 Reggio Emilia, Italy

Abstract

This study explores the application of topic modeling techniques for auditing purposes in the banking sector, focusing on the analysis of reviews of anti-money laundering alerts. We compare three topic modeling algorithms: Latent Dirichlet Allocation (LDA), Embedded Topic Model (ETM), and Product of Experts LDA (ProdLDA), using a dataset of 35,000 suspicious activity reports from an Italian bank. The models were evaluated using the coherence score, NPMI coherence, and topic diversity metrics. Our results show that ProdLDA consistently outperformed LDA and ETM, with the best performance achieved using 1-gram word embeddings. The study reveals distinct topics related to specific client activities, cross-border transactions, and high-risk business sectors, like gambling. These results demonstrate the potential of advanced topic modeling techniques in enhancing the efficiency and effectiveness of auditing processes in the banking sector, particularly in the analysis of activities that could be tied to money laundering and terrorism.

Keywords

Topic modeling, Auditing, Banking sector

1. Introduction

There has always been a close connection between banks and the collection of different kinds of empirical data: banks, just like any other company, have always poured large amounts of resources into understanding numbers, and how to deal with them. Numerical data, being closely related to the financial performances of companies, has always taken the spotlight.

On the other hand, linguistic data has always been much less considered, due to the difficulties of analysis and underwhelming performances.

But things are changing. More and more companies are understanding the value of language, which contains information that no number can convey. Different Natural Language Processing (NLP) tasks, language resources, and computational linguistics practices have now become a staple in many realities, like sentiment analysis [1] and word embeddings [2].

In fact, there is a wide variety of linguistic data that banks can exploit: emails, bank transfers descriptions, internal communications, and customer feedback. Some peculiar issues arise, when dealing with linguistic data in the banking sector, like the usage of acronyms, abbreviations and technical terminology. These data are often proprietary, meaning that the bank owns them, and the access is forbidden to externals. While the quantity of information they contain is massive, a downside is that the impossibility of sharing it with other banks hinders the possibility of a more global analysis.

In this context, this paper wants to explore the application of topic modeling techniques to the auditing process, in particular regarding the analysis of reviews of anti-money laundering (AML) alerts. Topic modeling can, in fact, be an incredibly helpful tool for auditors who want to perform an in-depth analysis on large amounts of data.

An overview of topic modeling algorithms and applications in the banking sector, both documented in scientific research and in concrete applications within banks, will be presented. Then, we will provide a comprehensive description of the data employed, followed by the preprocessing operations. We will then present the results and their interpretation, leading us into the conclusions. Finally, we will present a number of future works suggestions, which can expand this topic.

2. Related work

Topic Modeling is an unsupervised task of NLP, consisting in the extraction of latent themes in a given corpus. Latent Dirichlet Allocation, or LDA [3] is a probabilistic generical model, which became the most widely used and expanded-upon topic model. However, LDA faces several limitations, like scalability, low performances with large datasets, and the struggle against polysemy and homonymy [4].

To overcome the limitations of LDA, a lot of effort has been put into developing models that rely on word embeddings and neural networks, like ETM [5] and ProdLDA [6]. These models have been proved to provide better performances than LDA, at the cost of a higher computational effort[7].

In the last decade, topic Modeling has already been largely employed in the banking sector, and in auditing as well. [8] focused on the assessment and handling of frauds, while [9] analyzed financial misreportings. Another popular subject of analysis is accounting (for example [10]).

3. Data

The data employed is a collection of reviews of anti-money laundering alerts, that are automatically detected by a rule-based detection tool, whose name cannot be disclosed due to a specific request. This tool is widely employed across all Italian banks, and is aimed at tackling potential money laundering and terrorism financing schemes. It uses advanced algorithms to identify patterns that deviate from standard behavior.

An activity is considered suspicious whenever it exceeds certain risk thresholds. These activities are then reviewed by a human operator, who will evaluate whether the movement is actually tied to illegal operations or not. If the operation is not considered dangerous, or if there is not enough evidence to decide whether the activity is actually a threat or not, the operator will write a brief review, consisting of two sections. The first one is a description of the analyzed activity, The second section is either an explanation for why it was not considered dangerous; or a statement about the lack of evidence and the need to keep monitoring. This latter kind

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
alessandro.giaconia01@icatt.it (A. Giaconia); vchiariello@credem.it (V. Chiariello); sgiannuzzi@credem.it (S. Giannuzzi); marco.passarotti@unicatt.it (M. Passarotti)
0000-0002-9806-7187 (M. Passarotti)

Italian:
CASEIFICIO.MOVIM.COERENTE CON TIPO DI ATTIVITA'(ACCONTI A CONF.E PAGAM FORNITORI). IL CASEIF SI STA FONDENDO CON ALTRA LATTERIA, STA VENDENDO FORMAGGIO E SALDANDO I DEBITI.OK DOC REDD., OK ADEG.VERIF.NON SEGNALARE
English:
Cheese factory. Consistent movement withtype of activities (advance payments to contributors and payments to suppliers). The cheese factory is merging with another milk factory, it's selling cheese and settling debts. Income documentation is ok, adequate verification is ok. Do not report.
Italian:
TRATTASI DI FRUTTA E VERDURA ATTIVO SULLA PIAZZA DI ***UNICO FRUTTA E VERDURA DELLA PIZZA. ATTIVO CC CHE RACC INCASSI E ADDEBRELATIVI ALL'ATTIVITA'.AL MOMENTO NO PART ANOMALIE. MONITORIAMO
English:
Case of greengrocer active in the square of ***, only greengrocer in the square. Active bank account, that collects income and charges relative to the activity. No particular anomalies at the moment. We keep monitoring.

Table 1
Examples of sentences from the dataset with translations

of reviews usually ends with expressions such as 'monitoriamo' and 'continuiamo a monitorare'. The dataset employed consists of such reviews.

In Table 1 we provide two examples of documents, with their corresponding English translation. The English translations have been cleaned of abbreviations and spelling mistakes.

Due to hardware limitations, we worked using a selection of 35,000 documents, chosen randomly. The data is owned by Credem and is not publicly available, due to legal constraints. It is not possible to reveal the time period in which these documents where collected, nor the whole dataset size.

Each document has an average of 20.94 tokens per document.

It is important to note that the documents feature an abundance of spelling errors, abbreviations, acronyms, and missing blanks spaces between words. This in part due to a 300-characters limit. By comparing the tokens in the dataset with a dictionary of 4 millions Italian words[1], we obtain the results shown in Table 2:

Metric	Value
Total number of tokens	1,474,077
Total number of Out Of Vocabulary tokens (OOV)	193,482
Total number of OOV types	29,809
Number of sentences containing 1+ OOVs	60,870
Ratio of OOVs over the total number of tokens	0.1313

Table 2
OOVs in the complete dataset

The dictionary has been further enhanced in a data-driven approach, by including a list of Italian names[2] and surnames[3], and a list of the most frequent acronyms featured in the dataset, so that they are not incorrectly considered OOVs. In order to find the acronyms, we created a list of all OOVs in the dataset, in descending order, based on frequency. The 20 most frequent acronyms were added to dictionary, such as PEP (Persona Politicamente Esposta) and CC (Conto Corrente).

The table shows that about 13% of the dataset is made of OOVs. In comparison, the UD_Italian-ISDT treebank[4], tested against the same enhanced dictionary, contains only 6% of OOVs. For this comparison, the treebank in its entirety has been employed, consisting of training, testing and developing set.

The result shows a peculiar dataset, containing a considerable amount of OOVs, which will require robust methods of analysis.

Before processing the data, we performed data cleaning through stopwords removal and lemmatization.

Stopwords removal includes prepositions, articles, and conjunctions. This operation is helpful in reducing the number of tokens to be processed, gaining in efficiency, while also excluding data without semantic content. This operation was performed using the stopwords removal tool for Italian provided by Natural Language Toolkit[5] (NLTK).

After performing stopwords removal, the number of tokens in the complete dataset is reduced to 972,019, with an average of 13.47 tokens per document. Since we are using 35,000 rows, about half of the dataset, the number of tokens is 471,293.

Secondly, we performed lemmatization. The model employed is it_core_news_lg, provided by spaCy[6], which is made by 500.000, 300-dimensions-shaped vectors. Lemmatization is helpful in maintaining consistency through the whole dataset, as well as improving text understanding and efficiency. The spaCy model employed has a lemmatization accuracy of 97%, which is a satisfactory performance[7]. However, the model's performance on the dataset was tested. We created a sample of 100, randomly selected documents, who were then manually lemmatized, acting as the gold standard. The model's lemmas were then compared to the gold standard. The model's accuracy score was 79%, which is much lower than its usual accuracy. This underwhelming result further indicates how challenging to analyze the dataset is.

Before preprocessing, the TTR (Type/Token Ratio) was 0.0541; after this operation, the Lemma/Token Ratio is attested at 0.0428. The score is lower, indicating that we managed to reduce dispersion. Reducing dispersion is helpful in improving the performance of the algorithms, since word forms that used to be different are now considered to be the same.

[1] https://github.com/sigmasaur/AnagramSolver/blob/main/dictionary.txt
[2] https://gist.github.com/pdesterlich/2562329
[3] https://github.com/PaoloSarti/lista_cognomi_italiani/blob/master/cognomi.txt
[4] https://github.com/UniversalDependencies/UD_Italian-ISDT
[5] https://www.nltk.org/
[6] https://spacy.io/
[7] https://spacy.io/models/it

4. Processing

We have chosen three models for our analysis: LDA, ETM, and ProdLDA. These models were selected due to their different natures: the first is generative, the second is embedding-based, and the third is neural-network-based.

LDA assumes that each document is a mixture of topics and that each topic is a distribution over words. It uses Dirichlet priors to model the distribution of topics within documents and words within topics.

ETM represents words as vectors in a continuous space (word embeddings) and models topics as distributions over these embeddings, enabling it to capture more semantic relationships between words compared to traditional models like LDA.

ProdLDA is a neural-network based variant of LDA that uses a variational autoencoder (VAE) framework. ProdLDA models document-topic and topic-word distributions using neural networks, and it represents a "product of experts" model, focusing on improving topic coherence and overcoming the limitations of LDA.

The tool used for optimizing, training and comparing these models is the OCTIS (Optimizing and Comparing Topic Models is Simple!) library, developed by [11]. It allows users to compare the performance of various models with respect to different metrics, like Topic Diversity and Coherence Score.

Before training, a fundamental step is hyperparameters optimization, which controls the behavior of the algorithm, and therefore, its performance.

OCTIS allows to perform Multi-Objective Bayesian Optimization [12], a method that searches for the best hyperparameters configuration considering more evaluation metrics at once; in particular, the evaluation metrics we employ are:

- the Coherence Score, measuring how interpretable the topics are [13];
- the NPMI (Normalized Pointwise Mutual Information, measuring the statistical similarity of words inside a topic [14];
- Topic Diversity, measuring how different topics are from one another [15].

However, certain limitations need to be considered. In particular, the hardware employed was uncapable of handling such computational efforts; and, since the data is protected by privacy laws, using another, more powerful machine, is out of question.

To overcome this problem, we relied on SOBO (Single-Objective Bayesian Optimization)[16] which finds the best hyperparameters configuration with respect to only one metric. In particular, we chose the Coherence Score as the target evaluation metric. This metric was chosen due to its nature of measuring semantic coherence and, therefore, it can be considered a good indicator of topic quality. SOBO works by training the model n times, each with different hyperparameters. The output of this process is the configuration that provides the best result.

Algorithms were optimized and trained in four different configurations:

- without the enhancement of word embeddings;
- enhanced by 1-gram Word2Vec[17] embeddings;
- enhanced by 2-grams Word2Vec embeddings;
- enhanced by pre-trained embeddings.

The Word2Vec embeddings are created from our dataset. Table 4 shows the composition of these word embeddings.

We can check the quality of the created embeddings by employing the library Bokeh[8]. Bokeh allows us to perform interactive visualization, creating a representation of the vectorial space that can be easily examined. As we can see in Figure 1, the word embeddings create a plot where the different semantic fields are nicely divided and distinct from the others.

The pre-trained embeddings, instead, are trained on Common Crawl and Wikipedia[9]. The pre-trained embeddings composition can be seen in Table 5.

5. Results and discussion

In Table 6 we can find an average of the scores of the evaluation metrics for each model run, either enhanced or not enhanced by the aforementioned embeddings.

We can clearly see that ProdLDA provided the best performances across all runs. In particular, the dataset enhanced by 1-grams embeddings yielded the best overall performance, with an average score of 0.564. Much worse is the performance of both LDA and ETM, which failed at creating distinct and interpretable topics. In the reminder of this section, in Table 7 we show some of the topics created by 1-grams-ProdLDA, together with examples of the most relevant words associated.

The topics of 1-gram-ProdLDA were examined by seven bank employees, working in the auditing sector. They were then asked how interpretable the topics were, and to give a label, indicating what that topic was about. The chosen label for each topic was the most frequent one, assigned to that topic, by the employees. Out of the 12 topics created, only one was considered to be non-interpretable, confirming the excellent performance provided by ProdLDA. However, this non-interpretable topic was also the most frequent, as shown in Figure2.

We can clearly see the even distribution of the documents associated to each topic. The most frequent topic, labeled as "X", is the aforementioned non-interpretable topic, containing miscellaneous or difficult to categorize documents. Most of the topics refer to specific clients' activities, like bank transfers, payments, or activities related to the bank account.

There are also some more specific topics. An entire topic is dedicated to tobacconists and gambling. This kind of activity typically makes wide use of cash, which can potentially be tied to money laundering schemes. This level of specificity in auditing could indicate either regulatory requirements for these sectors or the bank's recognition of unique risks associated with these business types.

There is also a specific topic for suspicious activities with foreign countries or carried on by foreign users. Dealing with cross-borders regulations on transfers can be difficult for the bank, suggesting that particular effort should be put into developing efficient strategies for auditing cross-border activities.

Using 2-grams word embeddings was the best option for both LDA and ETM. However, in ProdLDA, 1-grams word embeddings provided a slightly better performance. Nonetheless, 2-grams were generally the better option, especially considering the sharp difference in ETM. On the other hand, enhancing the dataset with pre-trained embeddings did not result in a significant impact: the performance improvement of LDA was

[8] https://bokeh.org/
[9] https://fasttext.cc/docs/en/crawl-vectors.html

Model	Hyper-parameter	Values/[Range]
LDA	Num. of topics	`[2, 50]`
	α	`[0.001, 5]`
	β	`[0.001, 5]`
ProdLDA	Number of topics	`[2, 50]`
	Dropout	`[0, 0.95]`
	Num. of neurons	`100, 200, 300`
	Num. of layers	`1, 2, 3`
	Activation function	`softplus, relu, sigmoid`
ETM	Num. of topics	`[2, 50]`
	Dropout	`[0, 0.95]`
	Hidden size	`100, 200, 300`
	Activation function	`softplus, relu, sigmoid`

Table 3
Hyperparameters and values

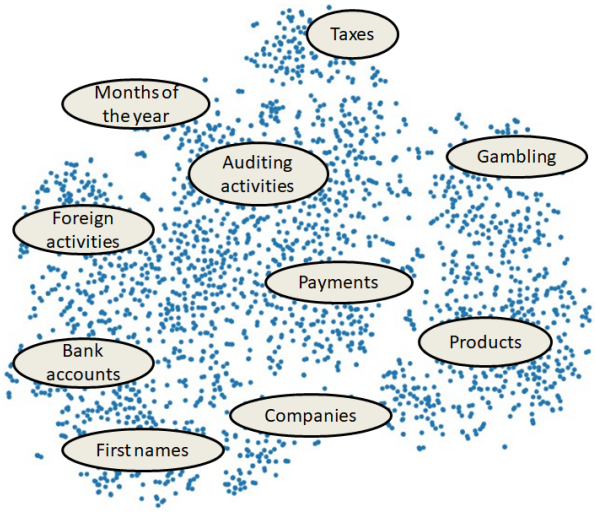

Figure 1: Vectorial distribution

minimal, while for ETM and ProdLDA it turned out to lower the outcome.

6. Conclusions and future work

NLP is now an essential component of the banking sector, and any company that wants to be competitive should make use of linguistic data science. In particular, in this paper we presented a NLP task, topic modeling, and how it can be imple-

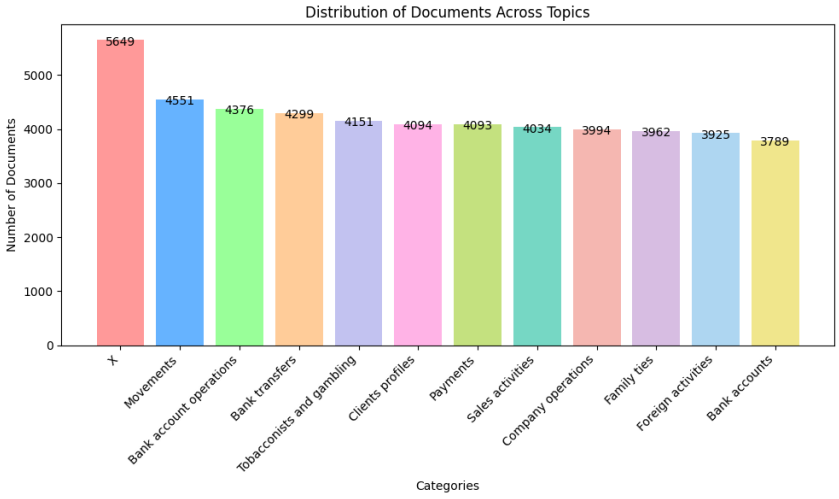

Figure 2: Topic distribution

Parameter	Value
min_count	20
window	5
vector_size	200
min_alpha	0.0007
number of negative samples	20
workers	6

Table 4
Word2Vec embeddings model parameters

Parameter	Value
Character n-grams	5
window	5
vector_size	300
number of negative samples	10

Table 5
Pre-trained embeddings model parameters

mented in the daily job of bank employees, in order to perform more detailed investigations. In particular, topic modeling can be a key component in the understanding and identification of money laundering schemes, as it allows auditors to perform more in-depth and focused analyses. For example, auditors could investigate patterns from the recent years, in order to have a better understanding on whether an activity is part of a larger trend, or an anomaly that deserves attention.

After citing other implementations of topic modeling in banking, we described the data employed, and its preprocessing, consisting in stopwords removal and lemmatization. Examples were provided, showing the peculiarities of the documents in the dataset. Then, the data was processed using three algorithms: LDA, ETM and ProdLDA. These algorithms were evaluated using three metrics: coherence score, NPMI score, and topic diversity. The optimal hyperparameters were found using SOBO. Optimization and processing were performed using four different configurations: without additional word embeddings, enhanced by 1-gram word embeddings created from our dataset, enhanced by 2-grams word embeddings created from our dataset, and enhanced by pre-trained word embeddings. The results show that ProdLDA's performance was far superior than its competition, especially when employing 1-gram Word2Vec embeddings. The algorithm outputted distinct and interpretable topics, which can provide a great insight into the data.

This experiment also has a large potential of being expanded. In particular, future works could employ a more computationally performing machine, in order to make use of the whole dataset, as well as performing MOBO, and obtain more precise hyperparameters. Finally, it is also possible to perform the same analysis on different kinds of data, in order to notice more clearly the differences and similarities from one kind of linguistic data to another, and their similarities. There are also new techniques that could have a great impact on this research, such as LLMs, Attention-based topic modeling, and Contrastive topic modeling.

	Embeddings				
	None	1-gram	2-gram	Pre-trained	Total avg
LDA	0.384	0.397	0.410	0.390	0.395
ETM	0.424	0.354	0.455	0.416	0.412
ProdLDA	0.552	0.564	0.552	0.535	0.550

Table 6
Average of the metrics' scores

Label	Top words
Tobacconists and gambling	tabaccheria bar lottomatica tabacchi servizi
Foreign activities	origine egitto periodo tunisia vacanza
Family ties	cointestato successione moglie fratello marito

Table 7
ProdLDA topics

References

[1] C. Nopp, A. Hanbury, Detecting risks in the banking system by sentiment analysis, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 591–600.

[2] I. Raicu, N. Boitout, R. Bologa, M. G. Sturza, Word embeddings in romanian for the retail banking domain, Bucharest University of Economic Studies (2020).

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[4] X.-Y. Jing, D. Zhang, Y.-Y. Tang, An improved lda approach, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34 (2004) 1942–1951.

[5] A. B. Dieng, F. J. Ruiz, D. M. Blei, The dynamic embedded topic model, arXiv preprint arXiv:1907.05545 (2019).

[6] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, arXiv preprint arXiv:1703.01488 (2017).

[7] X. Wu, T. Nguyen, A. T. Luu, A survey on neural topic models: methods, applications, and challenges, Artificial Intelligence Review 57 (2024) 18.

[8] M. Soltani, A. Kythreotis, A. Roshanpoor, Two decades of financial statement fraud detection literature review; combination of bibliometric analysis and topic modeling approach, Journal of Financial Crime 30 (2023) 1367–1388.

[9] N. C. Brown, R. M. Crowley, W. B. Elliott, What are you saying? using topic to detect financial misreporting, Journal of Accounting Research 58 (2020) 237–291.

[10] J.-C. Yen, T. Wang, A topic modeling-based review of digital transformation literature in accounting, in: Digital Transformation in Accounting and Auditing, Springer, 2024, pp. 105–118.

[11] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, A. Candelieri, Octis: Comparing and optimizing topic models is simple!, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 263–270.

[12] S. Terragni, E. Fersini, E. Fersini, M. Passarotti, V. Patti, Octis 2.0: Optimizing and comparing topic models in italian is even simpler!, in: CLiC-it, 2021.
[13] S. Syed, M. Spruit, Full-text or abstract? examining topic coherence scores using latent dirichlet allocation, in: 2017 IEEE International conference on data science and advanced analytics (DSAA), Ieee, 2017, pp. 165–174.
[14] S. M. Watford, R. G. Grashow, Y. Vanessa, R. A. Rudel, K. P. Friedman, M. T. Martin, Novel application of normalized pointwise mutual information (npmi) to mine biomedical literature for gene sets associated with disease: Use case in breast carcinogenesis, Computational Toxicology 7 (2018) 46–57.
[15] Y. Wu, X. Wang, W. Zhao, X. Lv, A novel topic clustering algorithm based on graph neural network for question topic diversity, Information Sciences 629 (2023) 685–702.
[16] P. Feliot, J. Bect, E. Vazquez, A bayesian approach to constrained single-and multi-objective optimization, Journal of Global Optimization 67 (2017) 97–133.
[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).