Topic Modeling: Reuters-21578 dataset

Rafael Soares
ISCTE-IUL
Mestrado em Engenharia Informática
Text Mining
Docentes: Fernando Batista, Ricardo Ribeiro

Tomás Machado
ISCTE-IUL
Mestrado em Engenharia Informática
Text Mining
Docentes: Fernando Batista, Ricardo Ribeiro

***Abstract*—Due to the high amount of digital sources of information that are released to the web each day, Natural Language Processing and Text Mining come to take advantage of this data in order to organize and categorize them for further purposes. This article studies the Reuters-21578 dataset and seeks for better understanding of different techniques and methodologies for topic modeling. To do so, various experiments were performed, with a variation on pre-processing methods, use of Bag-of-Words and TF-IDF for vectorization and use of Latent Semantic Analysis and Latent Dirichlet Allocation methods. In the article is concluded that different pre-processing methods can influence the outcome of the results, with the removal of stop words having high importance for the results. It was also concluded in the article that the use of TF-IDF for Vectorization and Latent Dirichlet Allocation brought better results for this kind of topic modeling.**

I. Introduction

We live in times where the use of the Internet is increasingly present in a daily basis of our lives, thus exponentially increasing the amount of data that is generated. The Internet made the access to information and news relatively easier for its users rather than the conventional way of checking the newspaper. Consequently, making the Internet growing as a way to access information and news, thus increasing the number of digital documents available. For this reason, Natural Language Processing (NLP), Text Mining and Machine Learning techniques are more and more popular as they can automatically organize and categorize documents.

In the scope of the Text Mining course, this article comes to further understand the functioning of topic modeling with the help of various Natural Language Processing and Text Mining techniques. To do so, this article explores the Reuters-21578 dataset, with a main focus on Latent Semantic Analysis and Latent Dirichlet Allocation methods to explore the topics covered in the dataset.

The article is divided in six main sections. In Section II, it is described work that was previously executed related to the study made in this article. Section III is related to the dataset description and all the different methodologies that were used for the execution of the experiments. A baseline was created and documented in section IV in order to better understand the LSA and LDA methods. Section V appears with the described experiments realized, with a variation in the use of methodologies described in section III. In Section VI all the results obtained from the experiments are described and discussed. Finally, section VII ends this article with the conclusions taken from this study and outlines possible future work.

II. Related Work

In the past years there were several articles created that seek to explain the importance of topic modeling and test different tools and techniques in order to obtain better results.

Zhang, Zhiwei, et al [1] conducted a topic modeling experiment using Latent Dirichlet Allocation (LDA) to model and discover topic structures in the Reuters-21578 dataset, with the main focus keeping the number of features low to see how it influences the categorization results. The underlying idea considered was that a good word for classification should be only relevant to one or few topics, rather than almost all topics. It was later concluded with their approach that it was possible to achieve better classification accuracy while reducing the feature space. Tao Liu made a similar experiment in [2] with the use of LDA for topic extraction, with the addition of a comparison with supervised methods. Reuters dataset was also used and analyzed during his experiment, concluding that the dataset was unbalanced due to the distribution in the categories. For this reason, only the top ten popular categories were used in his experiment.

In [3] it was made an experiment to compare the performance of three document representation methods, which are Term Frequency- Inverse Document Frequency (TF-IDF), Latent Semantic Indexing (LSI) and Multi-word. It was mentioned that the number of dimensions is a decisive factor for indexing and that LSI and multi-word have better semantic quality, and TF-IDF have better statistical quality. They also concluded that LSI showed better performance in categorization.

[4] approached with different methodologies over the Reuters dataset, with the addition of machine learning algorithms for classifying. In this paper, it is described that the dataset is divided into two, due to the fact that the dataset is unbalanced, where the first one only contains the top ten most frequent categories and the second one only contains documents associated with a single topic.

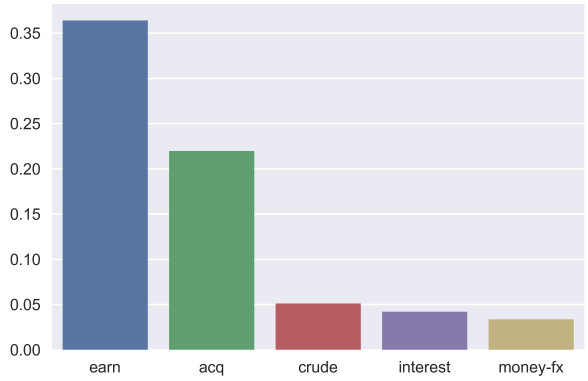

Fig. 1. Represents the normalized counts, for each of the first 5 most used topics.

III. Methodology

A. Data Description

For the elaboration of this paper it was explored the Apte-Mod version of the original Reuters-21578[1] dataset, which appeared on the Reuters newswire in 1987. The dataset contains a total amount of 10788 documents from the financial news, which are separated in a training set of 7769 documents and a test set of 3019 documents. Each document have a category associated with it (e.g., earn, interest, crude, money-fx, grain), with a total of 90 possible categories.

In Fig. 1 the Y axis corresponds to the percentage of documents and the X axis to a specific category. As we can see, the dataset is skewed due to the big difference in the percentage of documents from the most to the second most popular category, and then from the second to the third. There is also a high number of documents labeled with more than one category.

Since the dataset is highly skewed and the topic evaluation is made with unsupervised learning, without classification and validation, the results for the experiences further detailed in the article are evaluation in a contextual way.

B. Pre-Processing

This section will focus on the pre-processing methods used within the experiences realized in the following sections. The experiments are based on the single use and combination of the following methods in order to understand how they behave with the dataset and how to obtain better results:

- Transformation of all the words to lower case;
- Removal of Stop-words[2] obtained from a list obtained for the Reuters-21578 dataset;
- Removal of all words with length lesser than three characters, with the use of parsing.preprocessing.strip_short function from the Gensim library;

[1]https://github.com/teropa/nlp/tree/master/resources/corpora/reuters
[2]https://github.com/teropa/nlp/blob/master/resources/corpora/reuters/stopwords

Fig. 2. Simplified representation of the Methodology Pipeline used for our Work.

- Removal of all numeric characters with the parsing.preprocessing.strip_numeric function from the Gensim library;
- Part-of-Speech tagging (PoS): used to categorize each word from the documents into the corresponding Part-of-Speech [3]. The PoS[4] was applied from the Spacy Python library;
- Lemmatization: to group together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization[5] was applied from the Spacy python library.
- Name Entity Recognition (NER): to annotate named entities from documents. To do so, EntityRecognizer[6] from Spacy python library was used.

C. Vectorization

This section describes the vectorization methods performed after the pre-processing procedures.

- Bag-of-Words: Transformation of documents into a Bag-of-Words in order to obtain the frequency of each word per document, with the use of doc2bow function from the Gensim library;
- Term Frequency-Inverse Document Frequency (TF-IDF): to evaluate the importance of a word in a document and to convert the documents into a Vector Space Model. It was used TfidfModel[7] from the Gensim library for this implementation.

D. LSA and LDA Model application

This section describes the models applied in each experience after the pre-processing and Vectorization were executed.

Latent Semantic Analysis[8], also known as Latent Semantic Indexing is a Natural Language Processing (NLP) technique that analyses documents in order to find the underlying meaning or concept of these documents by performing a matrix decomposition on the term-document matrix.

Latent Dirichlet Allocation (LDA) is a probabilistic model with the purpose to learn the representation of a fixed number of topics and the distribution of the given topics in each document.

[3]https://en.wikipedia.org/wiki/Part_of_speech
[4]https://spacy.io/api/annotation#section-pos-tagging
[5]https://spacy.io/api/lemmatizer
[6]https://spacy.io/api/entityrecognizer
[7]https://radimrehurek.com/Gensim/models/tfidfmodel.html
[8]https://en.wikipedia.org/wiki/Latent_semantic_analysis

IV. Baseline

This section will focus on creating a baseline in order to have a comparison for further experiences in the construction of topic models . To do so, it will be focused on the application of Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

The following steps were realized in order to create a baseline:

- Simple pre-processing of the dataset, i.e, transformation of every word to lower case and removal of punctuation marks
- Convert all words from the dataset into a mapping between words and their ids per document using corpora.Dictionary[9] function from the Gensim Python library.
- Obtain the id for each word and the corresponding number of occurrences per document using doc2bow[10] function from the Gensim Python library
- Creation of a topic model using LDA and LSA methods
- Creation of a word cloud for the most important topics using both models
- Analysis of the models created through verification of most representative words in each topic and comparison with the original categories from the dataset
- Vary the number of topics and repeat the steps above, to understand the influence it has on the models

TABLE I
MOST IMPORTANT WORDS FOR EACH TOPIC USING 5 NUMBER OF TOPICS WITH LSA AND LDA MODELS

	LSA Model					LDA Model				
	w0	w1	w2	w3	w4	w0	w1	w2	w3	w4
t0	the	to	of	in	a	of	the	said	to	a
t1	mln	vs	dlrs	net	cts	the	in	to	a	of
t2	in	pct	vs	the	billion	the	to	of	in	said
t3	the	to	a	said	in	the	to	of	said	a
t4	nil	o	of	prev	wk	mln	vs	cts	net	dlrs

TABLE II
MOST IMPORTANT WORDS FOR EACH TOPIC USING 10 NUMBER OF TOPICS WITH LSA AND LDA MODELS

	LSA Model					LDA Model				
	w0	w1	w2	w3	w4	w0	w1	w2	w3	w4
t0	the	to	of	in	a	the	to	a	of	s
t1	mln	vs	dlrs	net	cts	the	in	to	of	pct
t2	in	pct	vs	the	billion	the	lt	to	of	and
t3	the	to	a	said	in	the	to	of	said	and
t4	nil	o	of	prev	wk	the	in	of	to	a
t5	to	of	s	the	u	the	of	to	said	and
t6	mln	vs	cts	pct	a	the	to	of	in	said
t7	s	dlrs	u	billion	to	vs	mln	net	cts	dlrs
t8	to	dlrs	billion	and	s	the	to	of	and	in
t9	of	to	said	a	and	cts	lt	the	april	to

The table I and II, with t0 to t9 representing the different topics and w0 to w4 representing the most frequent words for each topic, show the results obtained from each model, for the baseline experiences. After a brief analysis it is possible to conclude that the simple pre-processing, further pre-processing is necessary and the variation in the number of topics realized for the baseline experiences haven't brought much results, without any possible comparison from the topics captured by the models to the actual categories in the dataset.

V. Experiments

A. Experience 1

This experience was executed with the following pre-processing procedures:

- Transformation of every word to lower case
- Removal of all words with length lesser than three characters
- Removal of all numeric characters
- Removal of a list of Stopwords specifically made for the reuters dataset

For this experience it was used the Bag-of-Words for the Vectorization, followed by LDA and LSA models with 10 number of topics.

B. Experience 1.1

Experience 1.1 follows the same procedures as Experience 1, with a difference in the Vectorization, with this experience making use of TF-IDF as Vectorization from the SciKit-Learn[11] Python library.

C. Experience 2

This experience includes the same pre-processing, Vectorization and model application as Experience 1 with the inclusion of Lemmatization with Part-of-Speech Tagging, only considering names, adjectives, verbs and adjectives tags for the experience.

D. Experience 2.1

Experience 2.1 follows the same procedures as Experience 2, with a difference in the Vectorization, with this experience making use of TF-IDF as Vectorization, from the SciKit-Learn library.

E. Experience 3

This experience proceeds with the following pre-processing:

- Lemmatization;
- Part-of-Speech Tagging, keeping only words tagged as names, adjectives, verbs and adverbs;
- Name Entity Recognition (NER), keeping only words associated with an entity identified by this method;
- Removal of a list of Stop-words specifically made for the Reuters dataset.

For this experience it was used the Bag-of-Words for vectorization, followed by LDA and LSA models with 10 numbers of topics.

[9]https://radimrehurek.com/Gensim/corpora/dictionary.html

[10]https://kite.com/docs/python;Gensim.corpora.dictionary.Dictionary.doc2bow

[11]http://scikit-learn.org/stable/modules/feature_extraction.html

F. Experience 3.1

Experience 3.1 follows the same procedures as Experience 3, with a difference in the Vectorization, with this experience making use of TF-IDF as Vectorization, from the SciKit-Learn library.

G. Experience 4

This experience follows the same procedures as experience 3, but the documents where chunked into bi-grams.

H. Experience 4.1

This experience follows the same procedures as experience 3.1, but the documents where chunked into bi-grams.

VI. Results

This section describes the results obtained from the experiences referred in the previous section.

A. Experience 1

TABLE III
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LDA MODEL IN EXPERIENCE 1

	w0	w1	w2	w3	w4
t0	shares	lt	stock	company	bank
t1	production	prices	gold	crude	lt
t2	bank	january	february	year	market
t3	year	trade	government	dollar	growth
t4	tonnes	market	wheat	export	corn
t5	lt	april	record	group	shares
t6	dlrs	lt	loss	year	profit
t7	billion	marks	dlrs	february	surplus
t8	dlrs	billion	bank	week	banks
t9	trade	japan	agreement	bill	japanese

After taking a look at table III it is possible to see a huge difference from the baseline experience, only variating the pre-processing methods, adding a removal of numeric characters, words with lesser character length than three and removal of the Stopword list. It is possible to see some "it" words in the table, and this has happened due to a mistake made in the pre-processing phase, with an application of the strip_short function before the splitting of words. Another important thing to note is that it is possible to compare some of that topics to the actual categories labeled in the dataset, with a resemblance in the topic t4 with the categories "grain" or "corn". The other topics can also have a resemblance with the categories "earn","interest","trade","money-fx", due to the appearance of words in the table such as "shares","stock","production","growth", "profit".

B. Experience 1.1

After analyzing table III and IV it is possible to see a slight improve in the correlation between the words in each topic, with the only difference between each experiences being present in the Vectorization phase with the use of TF-IDF.

TABLE IV
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LDA MODEL IN EXPERIENCE 1.1

	w0	w1	w2	w3	w4
t0	loss	revs	profit	dlrs	year
t1	company	corp	shares	offer	dlrs
t2	tonnes	trade	year	production	japan
t3	mln	turnover	tax	billion	pre
t4	record	dividend	april	qtly	payout
t5	dlrs	shares	revs	corp	stake
t6	billion	dlrs	february	surplus	january
t7	qtly	april	prior	record	quarterly
t8	bank	dollar	rate	dealers	rates
t9	dlrs	earnings	quarter	share	year

TABLE V
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LDA MODEL IN EXPERIENCE 2

	w0	w1	w2	w3	w4
t0	share	dlrs	company	stock	april
t1	tonne	export	year	sugar	price
t2	cyclop	sale	corp	unit	company
t3	dlrs	year	rise	january	february
t4	company	corp	pacific	court	southern
t5	loss	dlrs	profit	revs	year
t6	bank	rate	market	dollar	currency
t7	trade	japan	oper	year	dlrs
t8	price	trade	country	export	stock
t9	company	share	offer	dlrs	analyst

C. Experience 2

After analyzing V it is possible to see that the pre-processing procedures bring similar results to the experience 1, showing some correlation between the words in each topic. The negative point about this experience is that it takes more time to execute the experience, so it is possible to conclude that it is not worth to implement the Lemmatization and Part-of-Speech methods.

D. Experience 2.1

TABLE VI
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LDA MODEL IN EXPERIENCE 2.1

	w0	w1	w2	w3	w4
t0	revs	dlrs	oper	loss	year
t1	unit	sell	corp	sale	company
t2	loss	profit	revs	sale	shrs
t3	dlrs	opec	steel	loan	store
t4	gold	tonne	wheat	department	program
t5	record	qtly	april	share	prior
t6	crude	price	dlrs	port	strike
t7	bank	trade	rate	market	dollar
t8	tonne	sugar	export	coffee	wheat
t9	dlrs	year	company	rise	february

In this experience a table wasn't elaborated any table, as the addition of TF-IDF have brought similar results to experience 2.

In this case, the addition of TF-IDF to the experience haven't brought much improve, with similar results to experience 2.

E. Experience 3

TABLE VII
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LDA MODEL IN EXPERIENCE 3

	w0	w1	w2	w3	w4
t0	bank	-	pct	rate	stock
t1	oil	u.s.	trade	barrel	import
t2	pct	year	rise	price	january
t3	mln	billion	dlrs	year	1986
t4	tonne	mln	sugar	u.s.	grain
t5	&	company	pct	corp	offer
t6	u.s.	trade	market	dollar	japan
t7	mln	&	share	ct	dlrs
t8	bank	gold	oil	price	mln
t9	mln	ct	loss	net	profit

TABLE VIII
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LSA MODEL IN EXPERIENCE 3

	w0	w1	w2	w3	w4
t0	mln	pct	year	dlrs	dlr
t1	mln	pct	ct	u.s.	loss
t2	pct	u.s.	trade	rise	billion
t3	share	mln	company	&	dlr
t4	loss	ct	bank	billion	mln
t5	billion	dlr	rate	bank	pct
t6	bank	loss	billion	ct	price
t7	oil	price	u.s.	trade	pct
t8	tonne	oil	billion	u.s.	stock
t9	year	ct	dlrs	price	billion

After taking a look at the tables VII and VIII it is possible to observe that LDA brings better results in comparison with the LSA model, as LDA has a more variety on the topics detected with a good correlation between words, even though the LSA model obtains good correlation between words in each topic, less topics are detected in comparison with the original categories in the dataset. It is also possible to see that this different experience have brought some good results, besides the unfortunate non alphanumeric characters showing as the most important words for some of the topics.

F. Experience 3.1

TABLE IX
MOST IMPORTANT WORDS OBTAINED FOR EACH TOPIC WITH LSA MODEL IN EXPERIENCE 3.1

	w0	w1	w2	w3	w4
t0	billion	feb	california	mln	merge
t1	ct	april	div	record	prior
t2	pct	trade	year	billion	rise
t3	mln	dlrs	quarter	loss	earning
t4	share	offer	company	lt	pct
t5	lt	company	corp	unit	acquire
t6	tonne	wheat	000	export	coffee
t7	bank	rate	oil	dollar	fed
t8	000	mln	ct	loss	net
t9	mln	stg	profit	gold	ct

In this case the use of TF-IDF brought better results, as it is possible to see more variety in the topics and that non alphanumeric characters don't appear as the most important words in comparison to experience 3. On the other hand, the word "it" and some numeric characters are appearing in this experience.

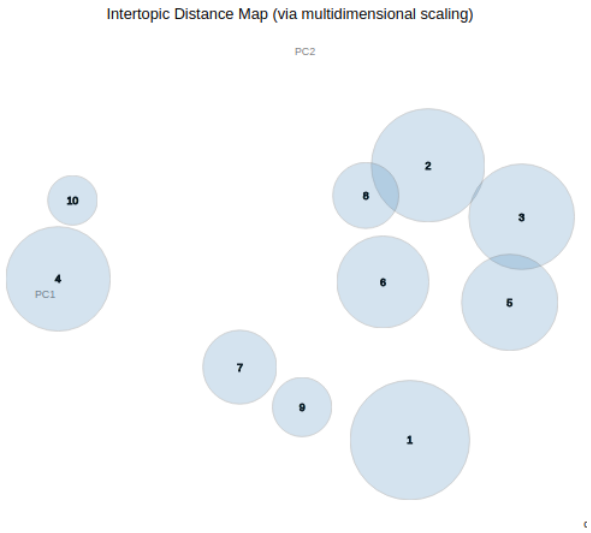

Fig. 3. Representation of the Isotropic Distance Map Dimensionality Reduction, from out Experience 2

G. Experience 4 and 4.1

Even though it was realized an experience with bi-gram chunks, the results weren't very promising in comparison with the effort taken to make this experiment.

VII. EVALUATION AND VISUALIZATION

In Machine Learning, evaluating a model can be done by having a preconceived understanding of what should come out of the Model. For this we had no previous understanding of, neither economics news articles or terms. Thus creating a limitation on the evaluations of the results. The author in [12], presents a survey from the research that has been done in the past 10 years in topic modeling evaluation and visualization.

For evaluating we mainly used our judgment, which was clearly not optimal, but, by our experiments was the best way to evaluate. **Coherence** metric was also used, but we could not find a baseline for this metric, as just changing the seed of the models would increase drastically the metric value, but not our evaluation of the topics.

pyLDAvis [6], was used to visualize the best models, but by reducing in our case 10 topics(dimensions) to 2 in order to visualize in a 2-dimensions plot is might not be a solution. As we present in fig 3.

A. Inference

For this part, we do the inference to the first 10 documents, that were not used for the training, and we only considered the first document label if the had more than one. By our

[12] github.com/mattilyra/pydataberlin2017/blob/master/notebook/EvaluatingUnsupervisedModels.ipynb

analysis we can see that some corpus news, have been in fact well categorized by ours topics, despite the subjectiveness of the news it self. Although this is impossible to quantify for this work, and will be discussed in future work. For example in our inference for the results of the Experience 2.1, the pre-processed inference document : *"exporters, fear, damage, japan, rift, mounting, trade, friction, japan, raised, fears, asia, exporting, nations, inflict, reaching, economic, damage, businessmen, officials, told, reuter, correspondents, capitals, move, japan, boost, protectionist, sentiment, lead, curbs, american, imports, products, exporters, conflict, hurt, long, run, short, term, tokyo," had a classification of* **trade, and was the corresponded topic from our model was the topic 7, which had the terms** *bank, trade, rate, market, dollar, japan, money, currency, deficit, reserve.*

VIII. Conclusion and Future Work

This article sought to analyze as much procedures as possible in order to understand how different methods behave with the studied dataset and the obtained results. It was concluded that the evaluations realized might not be 100% correct as they are measured by "eye balling methods" instead of being evaluated by a metric that can be compared. It is important to note the importance of the removal of Stop-words for this kind of classification and the use of the TF-IDF method with manipulated hyper-parameters. With this being said, it was considered that the best results obtained can be seen in table VI, with the use of pre-processing from Experience 1 and the addition of Lemmatization with Part-of-Speech tagging , use of TF-IDF for Vectorization and use of LDA model.

For future work, a deeper analysis of the dataset could be accomplished in order to better understand the vocabulary used in the documents and to understand the meaning behind each of the categories labeled in the dataset. Another example for future work would be to experiment different supervised machine learning techniques to avoid an analysis and evaluation based on "eye balling methods".

IX. Statement

- *Rafael Soares: 50%*
- *Tomás Manuel Machado: 50%*

References

[1] *Zhang, Zhiwei, et al. "An Efficient Feature Selection Using Hidden Topic in Text Categorization." 22nd International Conference on Advanced Information Networking and Applications - Workshops (Aina Workshops 2008), 2008.*

[2] *Liu, Tao. "Sparse Topic Model for Text Classification." 2013 International Conference on Machine Learning and Cybernetics, 2013.*

[3] *Zhang, Wen, et al. "A Comparative Study of TF*IDF, LSI and Multi-Words for Text Classification." Expert Systems with Applications, vol. 38, no. 3, 2011.*

[4] *Li, Tao, et al. "Efficient Multi-Way Text Categorization via Generalized Discriminant Analysis." Proceedings of the Twelfth International Conference on Information and Knowledge Management - CIKM 03, 2003*

[5] *Aseervatham, Sujeevan. "A Local Latent Semantic Analysis-Based Kernel for Document Similarities." 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.*

[6] *Sievert, C & Shirley, K.E.. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 63-70.*