Strategic Integration of Context for Fine-Tuning Topic Model Performance

Pierre Dardouillet, Kavé Salamatian, Hervé Verjus, Faiza Loukil, David Telisson, Olivier Le Van

Strategic Integration of Context for Fine-Tuning Topic Model Performance

Pierre Dardouillet*†, Kavé Salamatian*, Hervé Verjus*, Faiza Loukil*, David Telisson*, Olivier Le Van†

* LISTIC, University of Savoie Mont-Blanc
Annecy, France
Email: {firstname.surname}@univ-smb.fr

† Cegedim Business Services, Cegedim
Boulogne-Billancourt, France
Email: {firstname.surname}@cegedim.com

***Abstract*—Issue Tracking Systems software serves as an interface between a company and its customers. Customers can report bugs and seek assistance, among other demands. Reported issues include textual description, along with company defined metadata, aim at simplifying issue treatment by experts. In the context of the rapid growth of customer-reported issues, the manual treatment process becomes tedious and time-consuming. As a result, more and more studies focus on automating parts of this process, using semantic extraction and topic modeling approaches to automatically classify issues. To this end, most approaches consider the issue of textual description along with metadata, which can be a source of uncertainty and misleading in many real-world scenarios. Besides, knowledge from the company experts is often neglected. In this paper, we propose a general taxonomy of information incorporation into topic models. This aims to assemble all existing techniques, to further detect literature gaps. In addition, we propose a technique to incorporate expert knowledge into neural topic models. We evaluate our techniques and others in the literature on a real-world dataset coming from the JIRA software of a French HR management company. Results show a significant increase of more than 22% in classification performances when using expert knowledge, in addition to the issue textual description. The results validate our approach's effectiveness in improving the automatic classification of issues.**

***Index Terms*—Issue Tracking Systems; Neural Topic Models; Information Incorporation**

I. Introduction

Effective communication is the cornerstone for overcoming challenges and driving continuous improvement in the dynamic world of "customer engagement" interactions. Issue Tracking Systems (ITS) are vital components, that enable customers to report problems, seek assistance, and even propose improvements through a continuous agile project management approach. Prominent examples of ITS software include GitHub [1], Bugzilla [2], JIRA [3], among others. These platforms enable users to describe issues both in unstructured way, through plain-text fields, and structured way, through contextual tags like like severity, environment information, *e.g.*, software version, or other relevant labels. The structured elements enhance the quality of information given to expert troubleshooters, streamlining the issue resolution process. Even if ITS systems leverage on both structured and unstructured customers inputs, extracting relevant information from what is reported by customers is a daunting task. Recently, the focus has turned to the opportunities offered by artificial intelligence through advanced text processing and text semantic modeling. This entails analyzing data provided by customers for the automatic extraction of valuable and expert-interpretable insights, useful to improve the overall efficiency of the issue resolution process [4]. This is, nonetheless, a challenging endeavour, due to the sheer scale of the data, a big data issue, and the intrinsic complexity of information extraction from unstructured texts.

A promising approach in Natural Language Processing (NLP), relevant to this context, is topic modeling. It is designed to align, through an unsupervised learning approach, unlabeled text with labels or bag of keywords, describing the text. However, these labels should be pertinent to issue resolutions and point to different mitigation processes, leading the customer to the relevant unit within the company. If such labels can be identified, they will play a significant role in expediting issue resolution and reducing the time to solve them.

Several challenges exist on the path toward this objective. First, the inherent variability in data, even in structured contextual information, such as screenshots or predefined tags, may be misused because of customers' misunderstanding, resulting in a lack of the information needed by the troubleshooting experts. A second challenge concerns the unsupervised learning nature of topic modeling methods, leading to poorer classification performance than supervised one. Consequently, the research community has developed tools and methods to adapt topic modeling to semi-supervised approaches, using classification labels when available [5]–[7]. However, we decided not to consider such approaches as, in our case, predefined labels are generally not available. We are, therefore, in a typical data-mining context where we aim to enhance classification performance through leveraging contextual information present in various formats. A last and known limitation of topic modeling approaches is relative to their stability.

This property is crucial to trust the labels and topic extracted from inputs. However, the literature considering stability when evaluating models is scarce, and the lack of consensus on the metrics used for stability benchmark are almost inexistent [8].

The contributions of this study encompass multiple aspects, namely:

- A comparative study of topic modeling State-Of-The-Art (SOTA) methods, based on a novel dataset from a HR management system of a french company, extracted from JIRA. We propose a set of performance metrics that combine both expert requirements and insights from topic model. Results show the superior performances of neural topic models, compared to the classical LDA model [9].
- A novel taxonomy of information incorporation specific to topic modeling. The taxonomy encompasses all studies of the literature and classifies them following three dimensions: source, level, and incorporation technique. Our goal is to highlight topic model performance variations following each dimension and to uncover gaps in information incorporation techniques in the literature.
- A novel information incorporation technique using expert knowledge on topic priors is proposed. This technique is based on the topic seeds by Jagarlamudi et al. [10], which we adapt to neural topic models in this paper. Results from models with topic seeds defined by experts show the most significant increase in performances for classification and stability.

The ultimate goal is to identify best practices and techniques that significantly elevate the efficiency of the issue-tracking system, by automatically classifying ticket issues following expert-defined categories.

This article is organized as follows. Section II provides a review of existing topic modeling methods. Section III presents the taxonomy of information incorporation, along with the proposed topic seeds method for neural topic models. Section IV introduces the dataset used for experimentation, along with its challenges. Section V outlines the evaluation metrics. Finally, Section VI presents and discusses the obtained results.

II. Related Works

Topic Modeling, an application of NLP, has gained popularity over the past two decades. The underlying idea in topic modelling is that a given text is inspired by one or several topics, defining the keyword that will be used in the text. The goal of a topic model is to recover, from a given corpus of text, the latent topics that have inspired the words in the text, *i.e.*, each document of the corpus can be represented as a mixture discovered topics. Topics are represented by a word mix, the set containing the keywords attached to the topic [11].

A foundational topic model in NLP is Latent Dirichlet Allocation (LDA) [12]. LDA represents each document as a vector containing the observed word frequency in the document from an initial dataset vocabulary. LDA considers that documents are generated by a mixture of topics, where each topic is itself a mixture of words from the vocabulary. LDA assumes Dirichlet distributions priors for both the document-topic and topic-word distributions, characterized respectively, by hyper-parameters α and β. The statistical inference of the hyper-parameters is achieved in LDA using Variational Inference [13] and Gibbs Sampling [14], each contributing to the probabilistic assignment of words to topics and the iterative refinement of the model. The choice of hyper-parameters and the inference methodology significantly shapes the performance and stability of the LDA model.

A main limitation of basic LDA is its inability to benefit from metadata, and incorporating new modalities into it requires adapting the inference algorithm. Several extensions of LDA have been proposed to incorporate metadata, such as document's authors [15], and publication date [16]. Moreover, the Dirichlet mixture distributions used in LDA might not fit every use-cases. For instance, Correlated Topic Model [17], motivated by the possibility of certain topics appearing more often together, uses a log-normal distribution to model correlations between topics.

Over the past few years, neural topic models have rapidly grown as the new paradigms of topic modeling. They replace the inference problem in previous topic models with a more flexible optimisation problem. Variational AutoEncoder (VAE) architecture, first introduced by Miao et al. [18], has became popular neural topic model. Its architecture can be decomposed, as an encoder, learning the approximation of the reference distribution $p(\theta|X)$, where θ is the latent representation : $p(\theta|X) = q_e(\theta|X)$. Then, a decoder section uses this latent representation to discover a new document $X' = f_d(\theta)$.

Several variants of the VAE architecture have been proposed. In most of these variants, the encoder section f_e is implemented as a multi-layer perceptron, and the decoder section f_d utilizes a combination of a topic-term matrix β, with the latent representation: $f_d(\theta) = \theta * \beta$. The Neural Variational Document Model (NVDM) [18] employs a Gaussian distribution to model the variational distribution. Enhanced modeling performance is achieved by Srivastava et al. [19] through an approximation of the Dirichlet distribution using a Laplace approximation. They also adapt the decoder section to model the topic-word relations through a product of experts [20]. Another work is that of Liu et al. [21] accounting for correlation between topics by adding variables that learned during training.

A major advantage of neural topic models is their capability to incorporate metadata to assist in topic discovery. Bianchi et al. [22], [23] add sentence embeddings, through trained Transformers models like Sentence-BERT [24], to the encoder section. Embedded Topic Model (ETM) [25] model topic word relations using word embedding techniques like Word2Vec. Card et al. [26] incorporate various metadata attached to documents to assist topic extraction, document reconstruction, and prediction. This results into training the topic model in a semi-supervised way [27].

To sum up, topic modeling literature is broad and studies focusing on developing VAEs for topic modeling are widely represented in the literature. However, no study addresses extensively the information incorporation into topic models.

Some studies focus exclusively on pre-trained algorithms [22], while others on document metadata and labels [26]. To our knowledge, there is no comparison on how each incorporation type can affect the model behavior, and how this impact real applications.

III. Methodology

In this section, we introduce a taxonomy of information incorporation methods in topic models, covering existing studies in the literature. Subsequently, we then present expert information incorporation into VAEs, which to our knowledge, have not been addressed in the literature. Lastly, we propose a solution for quantifying stability in topic models, acknowledging that is a challenging task.

A. Taxonomy of Information Incorporation

We propose in Figure 1 a concise taxonomy of information incorporation in topic models. The aim of this taxonomy is to regroup existing studies in the literature, in a general framework, making it easier to identify similarities and differences between techniques.

We have identified threes axis relative to Information incorporation: *information source*, *information level* and *incorporation to topic model*. It is noteworthy that the term *information* is preferred to the term *data*, as it encompasses wider concepts.

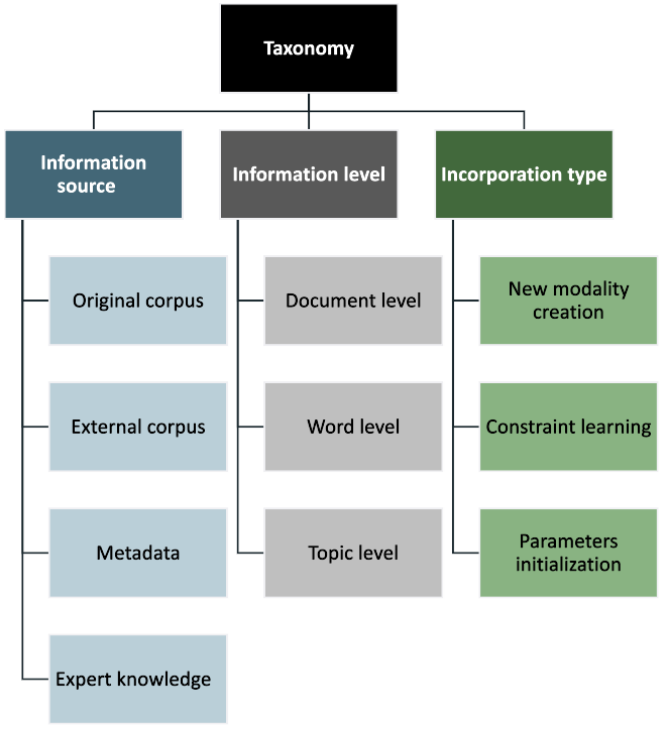

Fig. 1. A comprehensive taxonomy of information incorporation strategies.

- The **information source** axis represents the information origin. Information might have been extracted from the training document (also called training corpus), external corpus, or a set of metadata attached to the document. On other hand, information might come from other sources, such as general vocabulary, or expert knowledge specific to the application domain.
- The **information level** dimension refers to the granularity of additional information. Three case can be listed : document level, where new information are relative to the whole document, word level, and topic level. A given piece of information might belong to multiple categories, *e.g.*, document and topic level.
- The **incorporation** dimension considers the way of incorporating the new information into the model. Incorporation can be done through new modality, the information is processed and added to the topic model, in the encoder or decoder section. Alternatively, information can be incorporated as an optimisation constraint, *i.e.*, the new piece of information is used as regularization terms in the model training process. The third category is incorporation as parameter initialisation, *i.e.*, instead of setting randomly the initial value of the parameters, they are chosen based on incorporated information.

The proposed taxonomy can be used to categorize the literature in topic modeling studies. For example, for ETM [25] that uses word embeddings to model relations between words during the document reconstruction, the *information source* is the documents corpus, the incorporated information is at a word level, and words vectors are incorporated into the model by initialising internal parameters. For the supervised LDA [27] that uses document metadata (labels) to train the topic model in a semi-supervised manner, information source is document level metadata, that are incorporated into the topic model as an optimisation constraint.

A search in the literature reveals that a limited number of studies have utilized domain-specific knowledge, and topic-level information for VAE topic models. Nevertheless, existing studies, such as seeded LDA [10] show promising results in such incorporation to classical models, improving the clustering performance of models in various real-world applications [28], [29].

B. Seeded Topics

The aim of seeded topics is to leverage prior knowledge on the relations between words and topics, to guide the creation of topics around these words. The set of defined words per topic, referred to as S, takes the role of "anchors", influencing each topic to revolve around predefined words. Topic seeds, first introduced by Jagarlamudi et al. [10] for the LDA model, have undergone some adaptations to neural topic models [30], [31]. However, in this paper, we propose an incorporation of topic seeds into the neural model that does not require modifying the model structure, or objective function.

1) Seeded Initialisation: We use the simple approach for seed words per topic, to be incorporated in the topic-term matrix β of the model. The seeded matrix follows the following:

$$\beta_{i,j} = \begin{cases} V_+ & \text{if } \exists(\omega, t) \in S : \omega = i \text{ and } t = j \\ V_- & \text{if } \exists(\omega, t) \in S : \omega \neq i \text{ and } t = j \\ V_0 & \text{otherwise} \end{cases} \quad (1)$$

The role of the V_+ value is to heighten the importance of a term, for a specified topic. On the contrary, the V_- value decreases the importance of the same term for all other topics. Its purpose is to improve the diversity of topics, reducing the probability of a term in the seed to appear in other topics. Finally, the V_0 is a neutral value, used for all terms not existing in the seed.

2) Partial Seeds: In some cases, seeds are known only for a subset of the topics. While our method is not impacted by the lack of seed terms for a given number of topics, two distinct principles emerge:

1) We keep the Equation 1 unchanged. This means that the unseeded topics will have no V_+ values, but have V_- depending on other topics' seeds.
2) We limit the equation to affect only topics with known seeds. This means that unseeded topics will have only V_0 values, thus not discriminating terms that already appear in a seeded topic. The modified equation is then:

$$\beta_{i,j} = \begin{cases} V_+ & \text{if } \exists(\omega, t) \in S : \omega = i \text{ and } t = j \\ V_- & \text{if } \exists(\omega, t) \in S : \omega \neq i \text{ and } t = j, t \leq |S| \\ V_0 & \text{otherwise} \end{cases} \quad (2)$$

Where $|S|$ is the number of seeded topics.

In this paper, we only consider the principle in Equation 2.

3) Seed Values Determination: The choice for values V_+, V_- and V_0 is crucial when determining seeds. In neural topic models using the VAE architecture, the β matrix is unnormalised; meaning that values are not issued from a distribution, in the $[0, 1]$ range. Thus, β values are usually set to 0 at initialisation. To follow a natural distribution of values, we heuristically chose the seed values to be:

$$\begin{cases} V_0 & = 0, \text{ the neutral value} \\ V_+ & = a \in \mathbb{R}_{>0} \\ V_- & = \frac{a}{(|S|-1)} \end{cases} \quad (3)$$

The relation between V_+ and V_- assures that $mean(\beta) = 0$ at initialisation. The value of a is a variable, which impacts the importance of seed terms in their topics. We experiment with different values of a, and the results are reported in Section VI-B.

IV. Data Overview and Preprocessing

This section presents the dataset used for experiments, as well as the processing pipeline applied. Our dataset comes from the JIRA software and represents 10 years of issues concerning the management of administrative and human-related data of a company.

A. Dataset Description

Our dataset is composed of several issues. From each issue, we extract information generated at their creation, deemed useful for their classification. Below, we detail the extracted features of issues:

- The ticket's issue description, written by the issue reporter. This text is not subject to controls before the issue validation, and thus can be prone to typing errors, contain a mix of copy-pastes from software errors or codes, expert-only understandable values, and most importantly natural language. Removing all but natural language is a crucial task that we address in Section IV-B.
- The creation date of the issue. This metadata will not be used in further experiments, as issue topics have no links with their creation date, as per expert validation.
- The *Component* tag is selected from a predefined list by the user. These tags theoretically have a strong connection with the issue's topic, but in practice are not well used by the reporter. A study reported in Figure 2 shows that on a random subset of issues, most of the tags are set to a default value by the reporter. However, we consider this tag for further study, as it represents relatively pertinent metadata with uncertainty, as real-world applications usually contain.

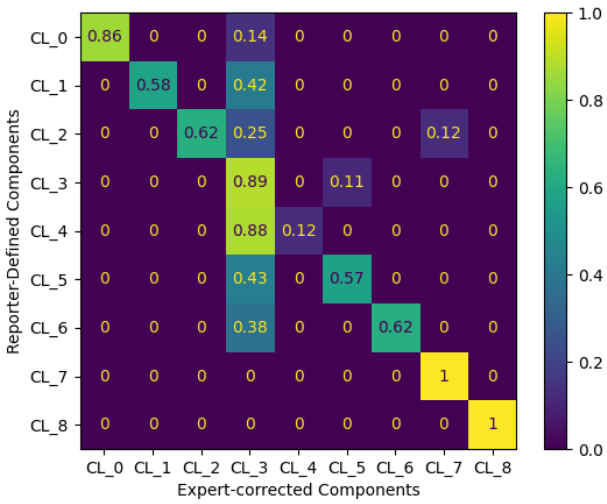

Fig. 2. Matrix representing the relations between the reporter-defined *components*, and the expert adjusted version. The 0.14 value of the first row indicates that 14% of *CL_0* defined by users are considered as *CL_3* by experts.

B. Dataset Processing

The document corpus, consisting of the description of the issues, first needs to be transformed to extract meaningful topics. The text processing techniques used can be separated into two axes. First are the classical techniques, comprising lemmatisation, tokenisation, and token filtering. Second are techniques specific to the dataset, such as name detection and removal for anonymisation, non-natural text deletion. The text

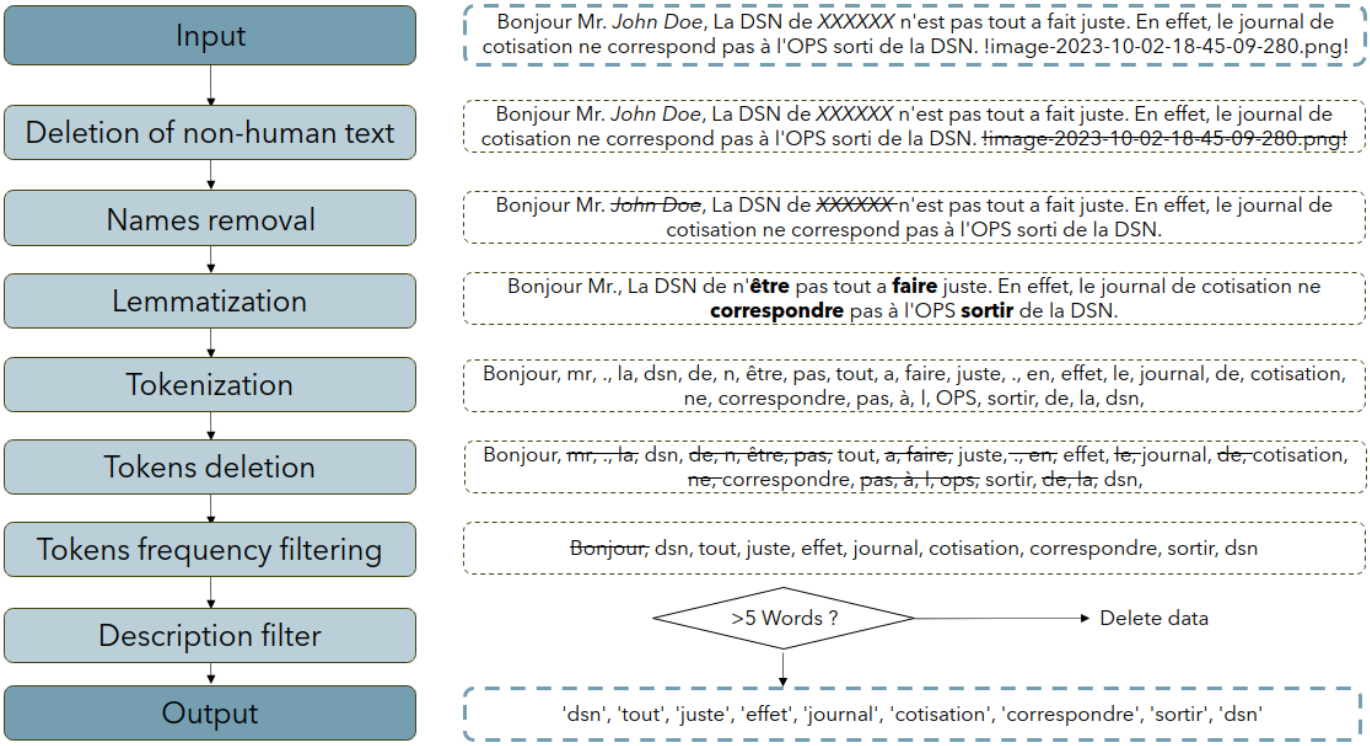

Fig. 3. Text preprocessing workflow with visual guidance.

processing pipeline is presented in Figure 3 and is described below:

1) **Non-human generated texts deletion:** Non-human texts, such as copy-pastes from code samples, error messages, or databases, should be removed from the descriptions. We use regular expressions to automatically detect the majority of non-human texts present inside markers, such as brackets {-}, [-], and other common symbols. Eventually, some non-human texts remain undetected, which remains in the final text.
2) **Names removal:** Texts are anonymised by detection and removal of names. To automate this task, we used a Named Entity Recognition (NER) model [32], which aims to detect named entities, such as names of persons, places, and organisations in texts. We applied the Camembert-NER [33] model, as it performs state-of-the-art in French texts. From detected entities, we only remove the person names.
3) **Lemmatisation:** To regroup words with the same meaning under a unique term, we use lemmatisation to transform all conjugations to their original form.
4) **Tokenisation** is then applied to transform the description to a list of words, which are used to constitute the vocabulary of the dataset.
5) **Tokens deletion** aims at removing words with no semantic bearing on a text. We delete short words (less than 3 characters), as well as French stopwords drawn from a list.
6) **Tokens Frequency filtering** helps reduce further the number of words in the dataset vocabulary, by deleting words not bearing topic-relevant meaning. We delete tokens present in more than 25% of the descriptions, as they are too common, and words present in less than 7 description, as they are too rare. Note that only two words were deleted from the $> 25\%$ filter: the French equivalent for *"Hello"* and *"Thank you"*.
7) At last, a **Description filter** is applied to remove descriptions having less that 5 tokens, after tokens filtering.

TABLE I
OVERVIEW OF THE EXTRACTED AND PROCESSED DATA.

	Extracted Data	Processed Data
Number of descriptions	4709	4501
Vocabulary size	17284	1576
Mean words per description	72.1	23.1
Min. words per description	2	5
Max. words per description	2339	449

Finally, our dataset contains 4501 issues and 1576 words in the vocabulary. Table I reports a summary showcasing key statistics, such as vocabulary size, mean, minimum, and maximum words per description, to illustrate the transformation and enhancement of the data during the preprocessing stage. Differences between vocabulary sizes and numbers of words per description are caused by the base Tokenisation algorithm, not filtering special and hidden character combinations, such as underscore lines "_ _ _ _ _", often used by reporters to format their message.

V. Evaluation Metrics

To evaluate the models' performances, we use a set of performance metrics, namely coherence, diversity, classification accuracy and f1-score, and stability. Each metric is defined below, and corresponds to different desired properties of a topic model, judged important for various applications. Coherence and diversity represent the topics' quality, in terms of relation between topics and words. Classification scores represent the model performances for downstream tasks. Finally, we quantify the stability of models, i.e. the difference between runs of the model with the same training corpus.

A. Analysis Metrics

The first category of measures focuses on the topic word connections. They represent the quality of topic model results for analysis tasks, where topic constitutions are primordial. Two complementary measures are considered namely, coherence and diversity.

1) Topic Coherence: Topic Coherence quantifies the interpretability of discovered topics from a human perspective [34]. Many coherence measures exist in the literature, mainly using Normalized Pointwise Mutual Information (NPMI) [35] and a sliding-window approach, over the corpus used for training the topic model. For our experiments, we use the C_v measure that is proposed in [34]. A C_v score close to 0.0 indicates low model coherence, thus low human interpretability, and a coherence score close to 1.0 indicates the best human interpretability.

2) Topic Diversity: Diversity measures the differences between topics discovered by a topic model. A higher diversity implies that most important words are different from one topic to another. High diversity is crucial in topic modeling, as it quantifies the model's discriminating power. To quantify diversity, we hereafter use the Rank Biased Overlap (RBO) [36].

B. Classification F1-score and Accuracy

Although both metrics presented above represent topic quality in an unsupervised way, this does not represent performances for downstream tasks, such as text classification. To measure this, we evaluate model classification performances using F1-score and accuracy based on a sample of expert-annotated tickets issues. The annotated dataset includes 9 different classes (User Interface, Time management, Accounting, Pay, Declarative documents, Manager software, Authorizations, Administrative Management, and Others) for 74 issues selected by an expert in the domain to represent a variety of issues present in the whole training dataset. The annotated dataset is separated from the training dataset, meaning that topic models see neither the issues nor the labels during training.

C. Stability Metrics

As topic models training are using stochastic algorithms, the produced topics can vary between different runs with identical hyper-parameters and training data, leading to instability of the results. Following different needs and intuitions, multiple methods to assess model stability are proposed in the literature [37]. In this section we motivate the stability measure we use in this paper and present its workflow in Figure 4.

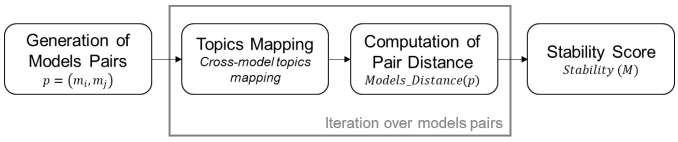

Fig. 4. Workflow of the proposed stability computation measure. First, we generate pairs of models. Then, for each pair, we determine the cross-model topics mapping, necessary to compute the models pair distance. Finally, we define the stability score as the mean model pairs distances, among all model pairs.

Our configuration for model stability computation considers multiple runs of the given topic model, using identical training data and hyper-parameters. We define the set of resulting models as $M = \{m_1, m_2, ..., m_r\}$, where r represents the number of models.

1) Generation of Models Pairs: The first step aims to generate all possible models pair among M, defined as $p = (m_i, m_j)$. As we use a symmetric distance measure on models pairs in the following steps, we consider the pairs (m_i, m_j) and (m_j, m_i) identical. Thus, the total number of pairs is equal to $|p| = r(r-1)/2$.

2) Topics Mapping: Stability relies on the distance between equivalent topics in two models. Therefore, for each models pair, we map topics from the first model, to those in the second. Following Hosseiny et al. [37] study, this can be achieved through two approaches: pairwise, and recurrent mapping. In recurrent mapping, multiple topics in a model can correspond to a single topic of the other model. In pairwise mapping, on the other hand, pairs of topics are mapped, meaning topics are related one to one. The latter approach provides a more natural representation of relations between topics, especially when considering topics independently, as in classification tasks. Consequently, we leverage pairwise mapping to assess the stability of the model, *i.e.*, we compute distances between each possible pair of topics in the two models. For this purpose two methods are investigated:

1) We use the topic-term matrix β to measure distances between words importance of a topic i from the first model, and topic j from the second model. The topic distance measure used is the Jaccard index, considering the set of the 10 words with the highest score for each topic, denoted $p_1(t_i)$ and $p_2(t_j)$ respectively.

$$Jaccard_p(t_i, t_j) = \frac{p_1(t_i) \cap p_2(t_j)}{p_1(t_i) \cup p_2(t_j)} \tag{4}$$

2) We use the topic mixture θ of the training corpus, and use the confusion matrix between topics of both considered models is used as a distance matrix between topics.

To identify the most suitable pairs of topics, we use a linear sum assignment algorithm [38]. The algorithm minimizes the

cumulative distance among the selected pairs, to match each topic of the first model to a topic from the second, assuring that each topic is matched only once.

3) Computation of Pairs Distance: Once topics are matched between both models, the models pair distance $Models\_Distance(p)$ can be measured as the mean of all topic pairs distances, $d(p_1(t_i), p_2(t_j))$, where $p_1(t_i)$ and $p_2(t_j)$ are two matched topics:

$$Models\_Distance(p) = \frac{1}{K} \sum_{(i,j)\in T} d(p_1(t_i), p_2(t_j)) \quad (5)$$

With K being the number of topics, d the topics distance measure and T is the set of matching topic pairs.

As in the first step, multiple topics distance measures d can be considered. The Jaccard index [39] and the Hellinger distance [40] are among the most used in stability quantification for topic models. The latter is more complete as it takes into account the importance of all words in the topic by considering their weights. However, it measures distance between probability distributions, and is not adapted to a range of neural topic models such as ProdLDA [19] or Scholar [26], that have topic-term matrix not issued from a probability distribution, unlike classical topic models. Thus, applying the Hellinger distance to these models is not straightforward.

4) Stability Score: Finally, similar to Belford et al. [41], the stability score is computed as the mean of model distances:

$$Stability(M) = \frac{1}{|p|} \sum_{p} Models\_Distance(p) \quad (6)$$

Where $|p|$ the number of models pairs depends of $|M|$. The normalization terms $1/|p|$ and $1/K$ from equation 5 allows the stability score to range between 0 and 1. A score closer to 1 indicates higher stability.

VI. Experimental Results

All the hereafter results are computed using an intel i7-1185G7 CPU, using python 3.10 and pytorch 2.1. For the Gibbs Sampling inference of LDA, we use the python library tomotopy[1] 0.12.5. For the variational inference, we use gensim[2] 4.2.0, with the $update\_every$ parameter set to 0 for batch learning. The results in this section represent the median score from 20 model runs.

A. Comparative study

Our first study is to compare the models' performances from the literature. Our goal is to lead a comparative study on our dataset, using only the description set available for each issue. We compare performances of conventional LDA model using Gibbs Sampling and Variational Inference, with neural topic models based on the VAE architecture. Results are reported in Table II. All models are trained with default parameters, and 9 topics to match the number of classes in the test dataset.

[1]https://github.com/bab2min/tomotopy
[2]https://github.com/piskvorky/gensim

Results show that neural models achieve overall better diversity scores than conventional models. However, other metrics show different results, where some neural models perform better than the classical LDA, and some perform worse. Overall, neural models based on the product of experts, namely ProdLDA, ContextualizedTM, and Scholar, achieve better coherence and accuracy than other models.

To gain a deeper understanding of the performances of the Scholar model, we run further tests with the model, varying the number of topics from 9 to 20. Considering our approach to foster classification performances and our limited test data, we chose not to explore the impact of the number of topics on the performances, as a higher number of topics leads to unrepresented topics on the test set, thus not being able to be coupled with a class. Results are shown in Figure 5.

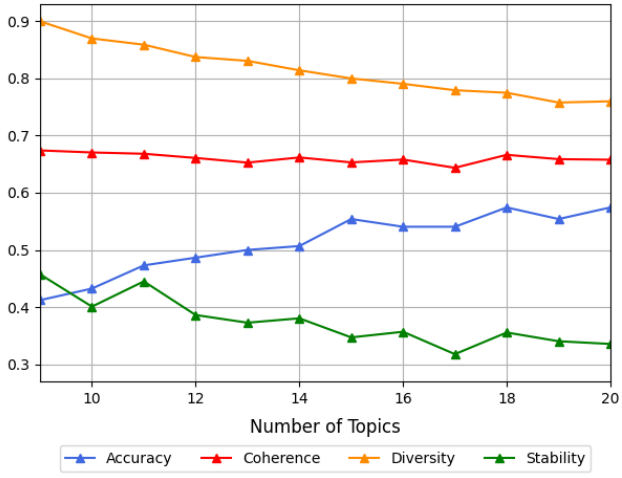

Fig. 5. Performances of Scholar per number of topics. The stability metric is computed from the topic-document matrix.

As the number of topics increases, classification scores get better, but stability and diversity decreases. When more topics are discovered by the model, the possibility of topic overlap increases, thus decreasing diversity. The number of internal parameters of the model also increases with the number of topics, thus causing a decrease in stability. As for classification scores, each topic represents sharper concepts, and multiple topics can correspond to a single class. Thus, the probability of a topic to encompass issues from different classes is reduced, inducing an increase in classification accuracy.

B. Information Incorporation Impact

To compare the information incorporation methods, we consider the Scholar and CombinedTM models with various added information, on multiple formats. Table III presents the main results, as variations of performances from the base model with no additional information.All results of this section are from models trained with 18 topics, as Figure 5 indicates it produces best classification performances.

Three experiments were conducted using the *Components* metadata as a covariate in the model: they correspond to a

TABLE II
PERFORMANCES METRICS OF EXISTING MODELS. THE PRESENTED RESULTS ARE MEDIAN SCORES OVER 20 RUNS OF MODEL TRAINING. THE BEST RESULTS ARE BOLD.

Model Type	Model	Analysis metrics		Classification metrics		Stability scores	
		Coherence	Diversity	Accuracy	F1-score	Topic-words	Topic-documents
Conventional	LDA Gibbs Sampling	0.61	0.82	0.38	0.37	0.27	0.40
	LDA Variational Inference	0.57	0.77	**0.41**	0.37	0.23	0.31
VAE	NVDM	0.39	0.97	0.27	0.26	0.04	0.01
	NeuralLDA	0.53	0.91	0.35	0.33	0.17	0.27
	ProdLDA	0.64	**0.98**	0.39	0.38	0.20	0.36
	ContextualizedTM Zero Shot	0.63	**0.98**	0.37	0.36	0.20	0.34
	ContextualizedTM Combined	0.63	**0.98**	**0.41**	0.38	0.22	0.38
	Scholar	**0.67**	0.90	**0.41**	**0.39**	**0.29**	**0.46**

model with only topic-word relations modeled (1), with the addition of topic-covariates relations (2), and with the addition of topics-covariates interactions (3). More details can be found in the Scholar original paper [26].

We remark that incorporating metadata as labels moderately improves classification performance, even if the proper classification head of the Scholar model is not considered here. Meanwhile, incorporation as covariates, shown in lines *"Components metadata as modality (1-3)"*, results in slightly inferior performances for most of the metrics. Two main factors can be the cause of these results: first, the metadata used is uncertain, thus biasing the model toward false relations between itself and the textual description. Second, as suggested by the authors of Scholar [26], the main contribution of covariate is the ability to model and investigate topics behaviors with variations of the covariate, such as sentiment, or date. This insight is not considered in this paper, as it is not the scope.

We highlight different behaviors for different types of information incorporation. Integrating information into the model through the initialisation of parameters consistently enhances model stability, whereas other incorporation techniques do not exhibit the same effect. This naturally demonstrates that replacing stochastic elements (mainly used at model initialisation) with external knowledge is the best method to increase stability. Moreover, all information decreases coherence, as the model does not only rely on text to discover topics, but takes into account new data modality, priors, etc. The coherence metric only takes into account the relation between the discovered topics and the text corpus.

However, the proposed taxonomy has some limitations. For instance, the simplicity of the taxonomy does not allow for representing differences in modality fusion techniques. However, the literature on this particular subject is vast [42], and different fusion types can lead to completely different behaviors from the model.

Using topic seeds to initialise the topic-term matrix demonstrates by far the best increase in classification accuracy and stability, paired with the worst performances in coherence and diversity. This illustrates that topic seeds chosen by experts, mainly motivated by issues classification, achieve the best results at shifting model behavior to favor classification performances rather than topic coherence, which is usually secondary when considering this task. A detailed study of topic seeds is presented in Figure 6, showing the impact of the parameter a on model performances.

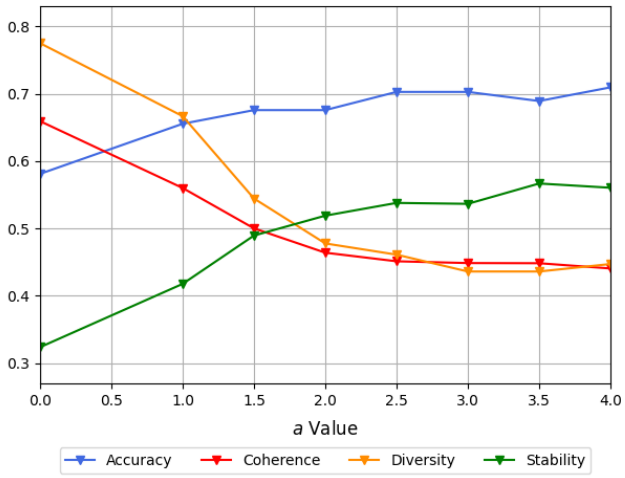

Fig. 6. Performances of topic seeds information incorporation in the function of a.

The a parameter from the Equation 3 can be seen as seed confidence: the higher the a, the higher the impact of seeds on the model. The topic seeds decrease the coherence and diversity of the topic model while increasing the accuracy and stability. The accuracy increases from 0.58 to a maximum of 0.70 with a value of $a = 2.5$. Similarly, stability increases from 0.32 to 0.54, coherence decreses by 0.21 points, and diversity decreases by 0.31. Increasing the a value beyond 2.5

TABLE III

IMPACT OF DIFFERENT TYPES OF INFORMATION INCORPORATION, FOLLOWING THE TAXONOMY PROPOSED IN SECTION III-A. RESULTS ARE DEVIATIONS (IN %) FROM THE BASELINE MODEL SCHOLAR WITH 18 TOPICS. THE PRESENTED RESULTS ARE MEDIAN SCORES OVER 20 RUNS OF MODEL TRAINING. IMPROVEMENTS FROM THE BASELINE ARE HIGHLIGHTED IN BOLD.

Description	Information			Analysis metrics		Classification metrics		Stability scores	
	Source	Level	Incorporation	Coherence	Diversity	Accuracy	F1-score	Topic-words	Topic-documents
Word2Vec word vectors on encoder	Training Corpus	Word	Model Init.	97,5%	94,3%	95,3%	97,0%	**115,7%**	**115,1%**
FastText word vectors on encoder	Training Corpus	Word	Model Init.	96,8%	88,5%	95,3%	97,1%	**115,4%**	**111,2%**
SentenceBERT pretrained model Camembert-base	External Corpus	Document	New Modality	93,5%	**113,3%**	89,4%	91,9%	84,3%	90,3%
SentenceBERT pretrained model Camembert-large	External Corpus	Document	New Modality	93,8%	**111,8%**	95,3%	98,4%	81,1%	86,6%
Components metadata as labels	Metadata	Document	Learning Algo	98,9%	**100,0%**	**101,2%**	**103,3%**	93,9%	91,1%
Components metadata as modality (1)	Metadata	Document	New Modality	98,8%	99,3%	94,1%	97,0%	96,1%	93,9%
Components metadata as modality (2)	Metadata	Document	New Modality	96,1%	**108,2%**	95,3%	95,9%	74,8%	74,2%
Components metadata as modality (3)	Metadata	Document	New Modality	89,7%	93,9%	92,9%	92,6%	89,1%	82,1%
Topic Seeds	Expert Knowledge	Topic, Word	Model Init.	66,1%	57,7%	**122,4%**	**124,7%**	**243,1%**	**154,5%**

has very little effect on model performances.

VII. CONCLUSION

In the context of customer reported issues, the need of automatic classification has grown significantly. In this scope, the application of topic modeling algorithms is used to explore interconnected textual data. However, limiting the focus to text alone falls short in capturing precise information of the issues in question. Incorporating various information to text documents is crucial to improve topic modeling applications in real-world scenarios. This information incorporation serves as a bridge, connecting machine-discovered topics with the desired properties identified by humans.

In this paper, we propose an in-depth study of information incorporation in neural topic models, taking advantage of the Variational AutoEncoder architecture. Our approach introduces a taxonomy that covers all facets of information incorporation found in the existing literature. Furthermore, we extend the topic seeds approach to incorporate expert knowledge on Variational AutoEncoder models. We lead a comparative study of 8 state-of-the-art topic models across various performance metrics, namely coherence, diversity, classification performances, and stability. This study uses a dataset from the JIRA issue tracking software concerning the HR management of a French company.

The obtained results underscore the remarkable performance of our topic seeds adaptation in downstream tasks, notably increasing significantly classification accuracy and stability, at the cost of a lower coherence and diversity. These results underline the practical effectiveness of our proposed model customer-reported issues classification.

The perspectives of this work involve conducting an in-depth study of topic seeds, considering the confidence parameter a to be specific to each word. Additionally, we aim to adapt the topic seeds to the model without a defined topic-term matrix. Furthermore, we want to investigate the classification performances of topic models for the highly specific ITS data, using a semi-automatic approach.

REFERENCES

[1] J. D. Blischak, E. R. Davenport, and G. Wilson, "A quick introduction to version control with git and github," *PLoS computational biology*, vol. 12, no. 1, p. e1004668, 2016.

[2] N. Serrano and I. Ciordia, "Bugzilla, itracker, and other bug trackers," *IEEE software*, vol. 22, no. 2, pp. 11–13, 2005.

[3] J. Fisher, D. Koning, and A. Ludwigsen, "Utilizing atlassian jira for large-scale software development management," Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2013.

[4] T. Diamantopoulos, N. Saoulidis, and A. Symeonidis, "Automated issue assignment using topic modelling on jira issue tracking data," *IET Software*, 2023.

[5] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine learning*, vol. 88, pp. 157–208, 2012.

[6] C. Zhang and H. Kjellström, "How to supervise topic models," in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13*. Springer, 2015, pp. 500–515.

[7] Y. Wang, Y. Yao, H. Tong, X. Huo, M. Li, F. Xu, and J. Lu, "Bug localization via supervised topic modeling," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 607–616.

[8] Y. Yang, S. Pan, Y. Song, J. Lu, M. Topkara, and J. Player, "Improving topic model stability for effective document exploration." in *IJCAI*, 2016, pp. 4223–4227.

[9] T.-N. Doan and T.-A. Hoang, "Benchmarking neural topic models: An empirical study," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4363–4368.

[10] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 204–213.

[11] X. Wu, T. Nguyen, and A. T. Luu, "A survey on neural topic models: Methods, applications, and challenges," 2023.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[13] M. Hoffman, F. Bach, and D. Blei, "Online learning for latent dirichlet allocation," *advances in neural information processing systems*, vol. 23, 2010.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl_1, pp. 5228–5235, 2004.

[15] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *arXiv preprint arXiv:1207.4169*, 2012.

[16] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[17] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[18] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*. PMLR, 2016, pp. 1727–1736.

[19] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[20] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[21] L. Liu, H. Huang, Y. Gao, Y. Zhang, and X. Wei, "Neural variational correlated topic modeling," in *The World Wide Web Conference*, 2019, pp. 1142–1152.

[22] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," *arXiv preprint arXiv:2004.03974*, 2020.

[23] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," *arXiv preprint arXiv:2004.07737*, 2020.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[25] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.

[26] D. Card, C. Tan, and N. A. Smith, "Neural models for documents with metadata," *arXiv preprint arXiv:1705.09296*, 2017.

[27] J. Mcauliffe and D. Blei, "Supervised topic models," *Advances in neural information processing systems*, vol. 20, 2007.

[28] L. Curini and V. Vignoli, "Committed moderates and uncommitted extremists: Ideological leaning and parties' narratives on military interventions in italy," *Foreign Policy Analysis*, vol. 17, no. 3, p. orab016, 2021.

[29] K. Watanabe and A. Baturo, "Seeded sequential lda: A semi-supervised algorithm for topic-specific analysis of sentences," *Social Science Computer Review*, p. 08944393231178605, 2023.

[30] W. Xu, X. Jiang, S. H. Sengamedu, F. Iannacci, and J. Zhao, "vontss: vmf based semi-supervised neural topic modeling with optimal transport," *arXiv preprint arXiv:2307.01226*, 2023.

[31] B. Harandizadeh, J. H. Priniski, and F. Morstatter, "Keyword assisted embedded topic model," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 372–380.

[32] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[33] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," *arXiv preprint arXiv:1911.03894*, 2019.

[34] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[35] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, vol. 30, pp. 31–40, 2009.

[36] S. Terragni, E. Fersini, and E. Messina, "Word embedding-based topic similarity measures," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2021, pp. 33–45.

[37] A. Hosseiny Marani and E. P. Baumer, "A review of stability in topic modeling: Metrics for assessing and techniques for improving stability," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–32, 2023.

[38] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.

[39] S. Pegg, "Dictionary of bioinformatics and computational biology," 2005.

[40] M. Daradkeh, W. Mansoor, S. Atalla, Y. Himeur, and O. Kerdjidj, "Lifelong machine learning for topic modeling based on hellinger distance," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 01–08.

[41] M. Belford, B. Mac Namee, and D. Greene, "Stability of topic modeling via matrix factorization," *Expert Systems with Applications*, vol. 91, pp. 159–169, 2018.

[42] F. Castanedo *et al.*, "A review of data fusion techniques," *The scientific world journal*, vol. 2013, 2013.