Semantically-Enhanced Topic Modeling

Felipe Viegas
UFMG - Brazil
frviegas@dcc.ufmg.br

Washington Luiz
UFMG - Brazil
washingtoncunha@dcc.ufmg.br

Christian Gomes
UFSJ - Brazil
christian@ufsj.edu.br

Amir Khatibi
UFMG - Brazil
amirkm@dcc.ufmg.br

Sérgio Canuto
IFG - Brazil
sergio.canuto@ifg.edu.br

Fernando Mourão
Seek AI Labs - Brazil
fernando.mourao@catho.com

Thiago Salles
UFMG - Brazil
tsalles@dcc.ufmg.br

Leonardo Rocha
UFSJ - Brazil
lcrocha@ufsj.edu.br

Marcos André Gonçalves
UFMG - Brazil
mgoncalv@dcc.ufmg.br

ABSTRACT

In this paper, we advance the state-of-the-art in topic modeling by means of the design and development of a novel (semi-formal) general topic modeling framework. The novel contributions of our solution include: (i) the introduction of new semantically-enhanced data representations for topic modeling based on pooling, and (ii) the proposal of a novel topic extraction strategy – ASToC – that solves the difficulty in representing topics in our semantically-enhanced information space. In our extensive experimentation evaluation, covering 12 datasets and 12 state-of-the-art baselines, totalizing 108 tests, we exceed (with a few ties) in almost 100 cases, with gains of more than 50% against the best baselines (achieving up to 80% against some runner-ups). We provide qualitative and quantitative statistical analyses of why our solutions work so well. Finally, we show that our method is able to improve document representation in automatic text classification.

KEYWORDS

Topic Modeling; Word Embeddings; Bag of Words

ACM Reference Format:
Felipe Viegas, Washington Luiz, Christian Gomes, Amir Khatibi, Sérgio Canuto, Fernando Mourão, Thiago Salles, Leonardo Rocha, and Marcos André Gonçalves. 2018. Semantically-Enhanced Topic Modeling. In *2018 ACM Conference on Information and Knowledge Management (CIKM'18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271797

1 INTRODUCTION

We live in a world of fast-paced and ever-increasing information. Properly representing such information, without loss, as well as developing effective and efficient strategies for handling and assessing it are paramount. Topic modeling is among the most exploited approaches to extract and organize information from large amounts of data. These approaches aimed at finding semantic topics from textual documents (e.g., product and app reviews, tweets). Such topics can be explored by downstream applications to accomplish a given task. However, traditional strategies for topic identification face a major challenge: the topics' semantics is usually extracted by just analyzing syntactic properties of textual data, assuming a high correlation between syntax and semantics.

In this paper, we advance the state-of-the-art in topic modeling, by means of the design and development of a (semi-formal) general non-probabilistic topic modeling framework. This framework is able to unlock useful information from textual data and, consequently, to uncover hidden "topics", being applicable to many domains. Our contributions also involve innovations in specific components of the framework. In more details, the proposed framework has three main building blocks, namely, (i) data representation (DR), (ii) latent topic decomposition (LTD) and (iii) topic extraction (TE). Briefly speaking, data representation (DR) has to do with how to encode the observed textual data in a proper structure that captures relevant information in a suitable manner. Several strategies do exist for this purpose, such as the traditional bag of words and the so-called word embeddings. In this work, we propose to exploit semantic-enhanced representations based on pooling to represent relevant information for the purpose of topic modeling. The LTD strategy has to do with the ability to decompose the properly represented dataset matrix into matrices that capture the latent relationships between terms and documents. Again, several strategies do exist for this matter, such as non-negative matrix factorization, LDA, and its probabilistic counterpart, to name a few. Finally, the topic extraction (TE) strategy has to do with how to extract the abstract "topics" from the factorized matrices produced in the previous stage. As we shall see, the introduction of a semantically-enhanced representation for the DR step brings a new problem: *how to properly represent topics within our solution, since there is no more a direct correspondence between topics and smaller semantic units (e.g. words) in our richer (Fisher-based) representations.* In this paper, we propose solutions for this problem as well.

To summarize, the main contributions of this paper are fourfold. *(i) the proposal of a general topic modeling framework able to discover more cohesive topics in textual datasets; (ii) the introduction*

CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6014-2/18/10...$15.00
https://doi.org/10.1145/3269206.3271797

of new semantically-enhanced data representations for topic modeling based on pooling; (iii) the proposal of a novel topic extraction strategy – ASToC – that solves the difficulty in representing topics in our semantically-enhanced information space; and, finally, (iv) a thorough evaluation of several instantiations of the framework, considering an extensive analysis of their strengths and drawbacks in terms of their main building blocks and interactions.

In our evaluation, we consider 12 datasets, comparing our solution against 12 state-of-the-art baselines for topic modeling with nine different evaluation metrics[1]. The best instantiation of our framework achieves the best results in the vast majority of cases. Particularly, we verify that the introduction of the aforementioned contributions (ii) and (iii) are key contributors to the observed gains against the second best baselines of more than 30% in coherence, 50% in mutual information and almost 80% in semantic similarity. Considering a universe of 108 experimental results (a combination of 3 evaluation metrics with 3 topic lengths over 12 datasets) we obtained the best results (99 best performances), with the runnerup baseline (GPU-DMM) coming far away, with just 30 top performances. We present qualitative analyses on the influence of the number of topics in our solutions, as well as the quantification on the impact of each framework building block in our final overall solution. We also demonstrate the suitability of our proposal for representing documents in automatic classification task, achieving improvements when compared with original representation and the GPU-DMM.

2 BACKGROUND AND RELATED WORK

2.1 Data Representation

The most traditional data representation strategy for textual documents is based on simple term occurrence information, encoded by the so-called TF-IDF score (and its variants). Although this approach is, by far, the most used one (especially considering learning approaches based on vector space models), it lacks useful information such as context. One simple strategy to overcome this is to use n-grams [5]. In the n-grams approach, the same importance weighting principle applies, but instead of considering single words, it considers a sequence of co-occurring words (or, simply, a context window). The use of n-grams has already shown to produce significant improvements in learning, although still limited in capturing contextual information observed in non-sequential patterns.

Recently, much has been developed in terms of data representations. Here, we pay special attention to the word embedding models, such as Word2Vec [18] and GloVe [21], which, based on co-occurrence statistics of textual datasets, represent words as vectors such that their similarities correlate with semantic relatedness. These strategies usually make use of contextual information, such as terms adjacent to a target term. Towards the design of a richer data representation, the authors in [18] propose the so-called Word2Vec model—probably the most popular word embedding strategy so far. Similarly to GloVe, the unsupervised Word2Vec strategy aims at estimating the probability of two words occurring close to each other. This is achieved by a neural network trained with sequences of words that co-occur within a window of fixed size, in order to predict the n-th word given words $[1, \ldots, n{-}1]$ or the other way around. The output is a matrix of word vectors or context vectors, respectively. Differently from other distributional models, both Word2Vec and GloVe are prediction models, in the sense that they aim at predicting word occurrence instead of only relying on co-occurrence patterns to represent data. This usually brings up richer representations that ultimately improve learning capabilities of downstream models. As shown in [1], prediction models consistently outperform count models in several tasks, such as concept categorization, synonyms detection, and semantic relatedness, providing strong evidence in favor of the superiority of word embedding models.

One common way of exploiting word embeddings for document representation is by pooling [15]. This proposal advocates that by properly pooling word vectors may lead to at least very competitive results. The authors evaluated the use of principal component analysis [15] (PCA) as an unsupervised pre-processing step that transforms the semantic vector space into independent semantic channels. After the pre-processing step, the authors propose the use of Fisher Vectors as a pooling strategy for word embeddings [15].

2.2 Latent Topic Decomposition

We now turn our attention to algorithms that aim at uncovering abstract topics from data. We start with the probabilistic models. In this case, data is modeled according to the following two main concepts: (i) each document follows a topic distribution $\theta^{(d)}$; and (ii) each topic follows a term distribution $\phi^{(z)}$.

Let $\theta^{(d)}{=}P_d(z)$ be the topic distribution over topics z observed in a document d. Also, let $\phi^{(z)}{=}P(w|z)$ be the probability distribution over terms w considering documents belonging to the abstract topic z. $\theta^{(d)}$ and $\phi^{(z)}$ reflect to what extent a term is important to a topic and to what extent a topic is important to a document, respectively. Both concepts are key for the so-called probabilistic latent semantic indexing (pLSI) [11], which takes into account the co-occurrence probability of terms and documents $P(d, w)$ as a combination of multinomial distributions conditionally independent.

Note that pLSI does not take into account how (ϕ) is generated, thus being harder to generalize to real-world scenarios. On the other hand, in [3] the authors propose the so-called latent Dirichlet allocation (LDA), which generalizes pLSI in terms of how $\phi^{(z)}$ is estimated: it assumes a Dirichlet distribution—a continuous multivariate distribution that reflects the fact that documents usually contribute to just a few topics and topics usually contribute to a small set of words. Despite being one of the most used strategies for latent topic decomposition, LDA has its own challenges, such as data sparsity and generation of incoherent topics. In [7], the authors proposed the bi-term topic model (BTM) method to deal with the data sparsity challenge, through the use of what they call bi-terms generated based on co-occurrence statistics of frequent terms. In [6], the authors deal with incoherent topics through a technique called lifelong topic model (LTM): an iterative method that exploits data from several application domains that usually show some degree of information overlapping in order to produce more coherent and reliable topics. The basic assumption here is that lexical and semantic relationships are key to uncover coherent topics.

In the basic pLSA, the word-topic distributions (Φ) and document-topic distributions (Θ) matrices are learned by directly optimizing the log-likelihood of the training dataset $L(\Phi, \Theta)$. In the recently

[1] Three metrics with three topic lengths.

developed Additive Regularization of Topic Models (ARTM) [24] approach, the basic pLSA model is augmented with additive regularizers. More specifically, the Φ and Θ matrices are learned by maximizing a linear combination of $L(\Phi,\Theta)$ and r regularizers $R_i(\Phi,\Theta), \forall i=1,\cdots,r$, with regularization coefficients τ_i as shown in Equation 1.

$$R(\Phi,\Theta)=\sum_{i=1}^{r}\tau_i R(\Phi,\Theta),\;\; L(\Phi,\Theta)+R(\Phi,\Theta)\rightarrow\max(\Phi,\Theta) \qquad (1)$$

Embedding-based Topic Model (ETM) [22] is another technique which incorporates the external word correlation knowledge into short texts to improve the coherence of topic modeling. ETM not only solves the problem of very limited word co-occurrence information by aggregating short texts into long pseudo-texts, but also utilizes a Markov Random Field regularized model that gives correlated words a better chance to be put into the same topic. *LDA, BTM, LTM, ARTM and ETM are used as baselines here.*

The FS method [10] is a strategy used to build topics with sentiment information. It extracts words that co-occur often (a.k.a. bi-grams). Then, it infers the sentiment strength of the extracted bi-grams based on the sentiment score of the documents in which they occurred. To generate the topics, the strategy applies LDA over these sentimental bi-grams. *We use FS as one of our baselines.*

We now consider the non-probabilistic topic modeling, comprising strategies such as matrix factorization. In this case, a dataset with n documents and m different terms is encoded as a design matrix $A\in\mathbb{R}^{n\times m}$ and the goal is to decompose A into sub-matrices that preserve some desired property or constraint. *Our proposed framework is specifically tailored for non-probabilistic strategies.*

A well-known matrix factorization applicable to topic modeling is the non-negative matrix factorization (NMF) [14]. Under this strategy, the design matrix A is decomposed into two sub-matrices $H\in\mathbb{R}^{n\times k}$ and $W\in\mathbb{R}^{k\times m}$, such that $A\approx H\times W$. In this notation, k denotes the number of latent factors (i.e., topics), H encodes the relationship between documents and topics, and W encodes the relationship between terms and topics. The restriction enforced by NMF is that all three matrices do not have any negative element. When dealing with properly represented textual data, the design matrix usually contains non-negative term scores, such as TF-IDF, with well-defined semantics (e.g., term frequency and rarity). It is natural to expect the extracted factors to be non-negative so that such semantics can be somehow preserved. *We thus consider NMF as our matrix factorization strategy of choice.* As a final note, as with the probabilistic strategies, the non-probabilistic ones can also generate incoherent topics, which is not desirable. We shall revisit this matter in next section.

Recent work have been proposed to improve the construction of topics by means of using word embeddings as auxiliary information for probabilistic topic modeling. Li et al. [16] propose a model called GPU-DMM, which can promote semantically related words using the information provided by the word embeddings within any topics. The GPU-DMM extends the Dirichlet Multinomial Mixture (DMM) model by incorporating the learned word relatedness from word embeddings through the generalized Pólya urn (GPU) model [16] in topic inferences. *GPU-DMM is one of our baselines.* To the best of our knowledge, there is no work that combines the information of word embeddings and non-probabilistic models. The main reason for the absence of works like ours is that the introduction of the richer information provided by word embeddings representation hampers the topics representation due to the lack of direct correspondence between topics and smaller semantic units (e.g. words) in these richer representations. As we shall see, we propose a new topic extraction strategy to mitigate this problem.

2.3 Topic Extraction

Once the abstract (latent) topics are uncovered, it is important to finally identify the relevant ones. As previously mentioned, both the probabilistic and non-probabilistic strategies for topic identification are prone to generate incoherent or irrelevant topics, which is undesirable. We now turn our attention to the techniques aimed at identifying relevant/coherent topics from previously identified topics.

The authors in [8] propose a supervised strategy called UTOPIAN. This strategy relies on human judgments to review the identified topics. It allows human agents to filter out incoherent topics, to merge distinct topics, to exclude terms from identified topics and to create new topics. For a fully automatic process, the authors in [9] propose the use of LSI followed by a clustering algorithm in order to identify relevant semantic topics.

Recently, in [2] a new technique, based on NMF, was proposed: the semantic topic combination (SToC). The main goal is to group semantically similar topics. Briefly speaking, after the NMF factorization of the design matrix A, a weighted tri-partite graph is built on top of matrices H and W (obtained from NMF factorization, $A\approx H\times W$), where the weights come from these sub-matrices. Such graph captures the relationship between documents and terms, as well as of terms and topics. A topic transition graph is then built through random walk on this tri-partite graph. Such random walk procedure is able to uncover indirect relationships between topics, ultimately grouping similar ones together, by means of the Bhattacharyya distance metric. At each interaction, the two topics with the smallest distance are merged, until all topics are finally merged. *This strategy, called BOW+NMF+SToC, is used as baseline here.*

3 PROPOSED STRATEGY

In this section, we present our proposed framework for topic modeling. The topic modeling is a well-known task, but to the best of our knowledge, *there is no previous work in the literature that formalizes the framework for non-probabilistic topic modeling*. We decompose our framework into three main building blocks, namely, Data Representation (DR), Latent Topic Decomposition (LTD) and Topic Extraction (TE). As we shall see, proper choices for each building block have a significant impact on the effectiveness of the topic modeling task. In this work, we go beyond the traditional Bag of Words (BoW), normally used in data representation, proposing the adoption of a richer semantic representation, Fisher Vector, described in Section 3.1. To the best of our knowledge, *it is the first time that Fisher Vector is used for non-probabilistic topic modeling tasks*. We use the NMF method in the LTD block in all analyzed variations of the proposed framework and, finally, three possible settings for TE block: (i) the absence of the block; (ii) the use of IG and (iii) ASToC. The IG and the absence of this block we will be detailed in Section 3.2. ASToC, our newly proposed method for TE, we will be explained in Section 3.3. **To summarize, we instantiate five viable combinations: (1) BoW + NMF; (2) BoW + NMF + IG; (3) BoW + NMF + ASToC; (4) FV + NMF + IG; and (5) FV + NMF + ASToC**. In Section 3.2, we explain why it is not possible to instantiate FV + NMF.

3.1 Fisher Vector Representation

As discussed in Section 2.1, documents are usually represented by a fixed-length vector representation (e.g., BoW with TF-IDF score), with dimensionality being the vocabulary size. Despite their simplicity and low computational cost, BoW representations usually miss contextual (or even semantic) information. One of the strongest trends aimed at overcoming this limitation is the use of word embeddings (e.g., Word2Vec[18], GloVe [21])—a vector space model representation that relies on contextual information in order to produce richer representations whose relative similarities between terms correlate with their semantic relatedness.

As we shall detail, we here represent words as vectors, which can be learned by means of strategies such as Word2Vec or GloVe. We represent each document as a multi-set of word vectors, where the order and the number of occurrences of the same word vectors do not affect the final representation. Recall that for non-probabilistic topic models it is key to factorize a design matrix $A\in\mathbb{R}^{n\times m}$, where n denotes the number of documents and m the number of features representing documents. In order to apply a decomposition method over A, it is important to transform each multi-set of word vectors (representing documents) into fixed-length vectors. Some strategies do exist for this purpose, such as averaging or max-polling. We adopt an already proven effective polling strategy called the Fisher Vector (FV) of single multivariate Gaussian distribution.

More specifically, consider a collection of n documents d_i ($1\leq i\leq n$) initially represented as a multi-set of word vectors, $W_i|_{i=1}^{n}=\{\overrightarrow{w}_1,\overrightarrow{w}_2,\cdots,\overrightarrow{w}_N\}$, where $\overrightarrow{w}_j\in\mathbb{R}^D$, $1\leq j\leq N$ and N is the number of word vectors in the multi-set W_i. A FV of a single multivariate Gaussian is a simplified version of the standard FV [15], defined as the gradient of the log-likelihood of W_i with respect to the parameters of a pre-trained diagonal co-variance Gaussian mixture model. In this case, there are no latent variables and it is possible to estimate the parameters $\lambda_i=[\overrightarrow{\mu_i},\overrightarrow{\sigma_i}]$ of this single diagonal co-variance Gaussian through maximum likelihood derivations, $\mathcal{L}(W_i|\lambda_i)$. Here, $\overrightarrow{\mu_i}\in\mathbb{R}^D$ and $\overrightarrow{\sigma_i}\in\mathbb{R}^D$ are estimated using the corresponding word vector in $W_i(1\leq i\leq n)$. The simplified FV is built from the gradient vectors given by the following partial derivatives, $\frac{\partial\mathcal{L}(W_i|\lambda_i)}{\partial\mu_i}=\sum_{j=1}^{N}\frac{w_j-\mu_i}{\sigma_i^2}$, $\frac{\partial\mathcal{L}(W_i|\lambda_i)}{\partial\sigma_i}=\sum_{j=1}^{N}\left(\frac{(w_j-\mu_i)^2}{\sigma_i^3}-\frac{1}{\sigma_i}\right)$, where $w_j\in W_i$.

In the standard FV, the diagonal of the Fisher information matrix F is approximated in order to normalize the dynamic range of the different gradient vectors' dimensions. Since we are using a single Gaussian model, the terms of the approximated diagonal Fisher information matrix become $F_{\mu_i}=\frac{N}{\sigma_i^2}$ and $F_{\sigma_i}=\frac{2N}{\sigma_i^2}$. Finally, the FV $\overrightarrow{F}_i$ of a document d_i is the concatenation of the two partial derivative vectors $\overrightarrow{V_\mu}$ and $\overrightarrow{V_\sigma}$, where $V_{\mu_i}=F_{\mu_i}^{-\frac{1}{2}}\frac{\partial\mathcal{L}(W|\lambda)}{\partial\mu_i}$ and $V_{\sigma_i}=F_{\sigma_i}^{-\frac{1}{2}}\frac{\partial\mathcal{L}(W|\lambda)}{\partial\sigma_i}$.

3.2 Latent Topic Decomposition

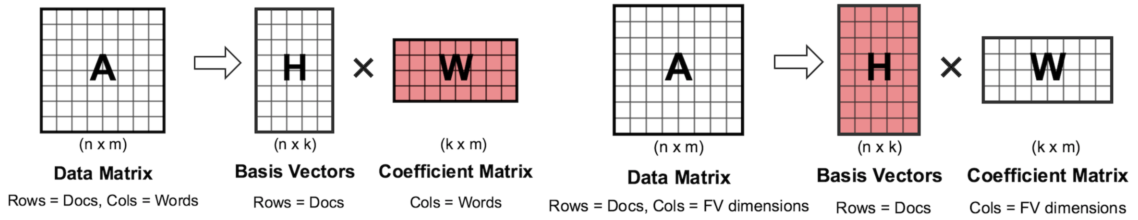
Non-negative matrix factorization (NMF) is an unsupervised family of algorithms that simultaneously perform dimension reduction and clustering, with successful applications in a wide range of domains, including topic modeling.

NMF [14] produces a "part-based" decomposition of latent relationships of a non-negative design matrix $A\in\mathbb{R}^{n\times m}$, where n is the number of examples and m the number of words. The goal is to find a k-dimensional approximation of A ($k<m$) in terms of non-negative factors $H\in\mathbb{R}^{n\times k}$ and $W\in\mathbb{R}^{k\times m}$. As we can see in Figure 1-a, H encodes the relationship between documents and topics while W encodes the relationship between terms and topics. The key idea behind NMF is to approximate the column vectors of A by non-negative linear combinations of non-negative basis vectors (columns of H) and the coefficients given by the columns of W. However, defining k for NMF is not a simple task since it is widely believed that NMF is a non-convex problem with a unique solution with no guarantee of finding the global minimum. The most usual way to extract the top words to represent a topic is done through the residual matrix W (Figure 1-a), since, the matrix W has the information about the words of the collection. Thus, for each topic i (i-th line W), the x words of greatest latent value are selected to represent the topic i, where x is the number of words selected to represent a topic. Using only the information of matrix W corresponds to the instantiation (3) BoW + NMF in our framework. As for the FV representation (Figure 1-b), the residual matrix W becomes the dimensions of the Fisher vectors. NMF loses the information about the words in the matrices decomposition and that is why we cannot instantiate FV + NMF. Another way to select the top words of a topic would be indirectly by using matrix H. Therefore, for each topic i (i-th column of H), we cluster the documents with no null latent value and then, apply a feature selection (i.e., IG) to rank the words (FV + NMF + IG). With the ranked words, we can select the top ones to represent topic i. For instance, for a topic t, the documents with no null values in H are $\{d_1,d_3,d_5\}$. Then, we can apply the Information Gain method over these three documents to find the top x words.

Using the information of H and IG is a way out to select the top words for each topic. However, we believe this solution may be less informative than the selection by W. This assumption may be true for both representations. For BoW, this is obvious as NMF generates a matrix that represents topics as words. As for the FV representation, this happens because both matrices (H and W) carry information about documents instead of words. In Section 3.3, we present a method that overcomes this issue.

3.3 Topic Extraction

In [2], the authors argue that factorization matrix strategies (e.g., NMF) might produce semantic sub-topics, but with no guarantees that the latent factors represent distinct and cohesive semantic topics. To address this issue, the authors propose a strategy, named SToC, for the hierarchical representation of semantic topics of a dataset. More specifically, given the residual matrices of a NMF decomposition, SToC determines which latent factors should be merged. We here propose an extension of SToC, called, Advanced Semantic Topic Combination (ASToC), designed to improve the cohesion of the topics extracted from the NMF method. The intuition behind the ASToC method is to use the information of both matrices (H and W) to merge topics and more importantly, enrich the information for extracting the top words for each topic. We consider StoC as a suitable choice for the TE block because of its potential to correlate documents to 'generic features', though the original method has only be used with BoW. ASToC builds on top of StoC by introducing three important extensions: *(i) a clear criterion to stop*



(a) BoW representation

(b) FV representation

Figure 1: Non-negative matrix factorization decomposition

topic merging; (ii) introduction of strategies to group documents represented as Fisher Vectors[2] *into their corresponding topics; (iii) the use of IG for selecting the best words for each topic.* As we shall see, ASToC is an excellent strategy for the TE block, especially when combined with the FV representation. It enhances the information provided by NMF, thus making it possible to select words for each topic using richer information than the residual matrix W, which is further filtered out using IG. The ASToC strategy is depicted in Fig. 2.

The key aspects are as follows. First, a weighted tripartite graph G_t is built. G_t has three types of nodes, namely, D, L and F, representing the documents, latent factors and features, respectively. Each weighted edge represents the intensity of the relationship between two nodes. The weights are given by a probability distribution derived from the information of the residual matrices of the design matrix decomposition. The values of H represent the edges $d{\leftrightarrow}l$ while the values of W represent the edges $l{\leftrightarrow}f$, where $d{\in}D$, $l{\in}L$ and $f{\in}F$. The intuition is that a document may refer to one or more topics (latent factors), whereas a feature can be associated with more than one topic (latent factor). Then, ASToC is applied to convert the tripartite graph G_t into a new graph G, describing only the relationships among latent factors (nodes L). A key aspect to be observed is that each latent factor $l_j{\in}L$ in G_t can be indirectly reached by each latent factor $l_i{\in}L$, thought nodes $f{\in}F$ and $d{\in}D$. These indirect links can be converted into direct ones according to

$$P(l_i{\to}l_j){=}\sum_{k\in D^{l_i}} P(d_k|l_i){\times}P(d_k|l_j){+}\sum_{k\in F^{l_i}} P(f_k|l_i){\times}P(f_k|l_j). \quad (2)$$

Each link of G summarizes this 2-step path of G_t, where F^{l_i} and D^{l_i} are sets of features and documents with an input edge from l_i.

The resulting graph G is represented as a stochastic matrix M, where the sum of each row $i{\in}M$ must be equal to 1. Next, the semantic sub-topics are merged in order to increase topic cohesion. The key idea here is to merge pairs of topics with a mutual probability of reaching each other in G higher than the minimum transition probability of G. To this end, we use the iterative algorithm, whose input is the graph M. At each iteration, the minimum transition probability P_{min} is computed. P$_{min}$ is defined among all pair of topics i,j in M as the mean probability of a random walk leaves i times the probability of a random walk goes from i to any topic chosen at random. Then, for each pair of topics i and j, we compute the cohesion of merging i and j into a new topic using the Equation 3, where cohesion is measured by the mutual transition computed by means of the Bhattacharyya distance, which penalizes pairs of topics with unbalanced or low probability of reaching each other.

$$cohesion{=}\sqrt{M[i,j]{\times}M[j,i]}{\cdot}(1{+}|Merge(i,j))|^{-1}) \quad (3)$$

The probability of i and j is normalized by the inverse of the number of distinct latent factors that compose the new topic. This normalization penalizes merging big topics since they usually have high input transition. Finally, we define the impact (Equation 4) of merging two topics through the cohesion (mutual transition probability) normalized by the minimum transition probability. The pair of topics with the highest impact is selected to be merged. In this case, both topics are removed from G while the brand new topic is added to it.

$$impact(i,j){=}\frac{cohesion{-}P_{min}}{P_{min}} \quad (4)$$

This process continues until the method reaches the final number of k topics. In our experiments, we empirically discovered that the best configuration for this parameter is a reduction in about 50% of the topics discovered by the NMF. In other words, NMF produces $2{\times}k$ topics which are merged according to the described procedure producing k topics as the final output of the TE block.

The next step is to group documents into their corresponding topics. Two operations are critical for this purpose, the first one being to update properly the document-topic matrix H produced by NMF, and the actual extraction of topics from documents. Recall that each column of H represents a topic. Such matrix must be properly updated to reflect the merging procedures just applied. We do so by combining columns corresponding to merged topics into a single column that represents the newly created one. If topics i and j were merged together, then we combine columns H_i and H_j into a single column. This may be accomplished in several distinct ways. We here adopt a very simple averaging strategy, where the new column $H_\bullet{=}\frac{H_i+H_j}{2}$. In this case, columns H_i and H_j are removed from H and $H_\bullet$ is appended to H as the last column. As this process finishes, we extract the relevant topics from documents. To this end, we consider documents grouped according to the discovered latent topics, considering the relevance of documents per topic represented in matrix H. Finally, for each set of documents belonging to a latent topic, we apply Information Gain [25] to rank terms according to their importance to the latent topics. The top x best-ranked terms from latent topic T are then selected as a relevant extracted topic.

4 EXPERIMENTAL EVALUATION

In this section, we analyze the proposed solution, comparing it against a large number of baselines and considering many datasets under a representative set of evaluation metrics.

[2] In fact, ASToC can group documents into topics for any type of document vector representation, including BoW.

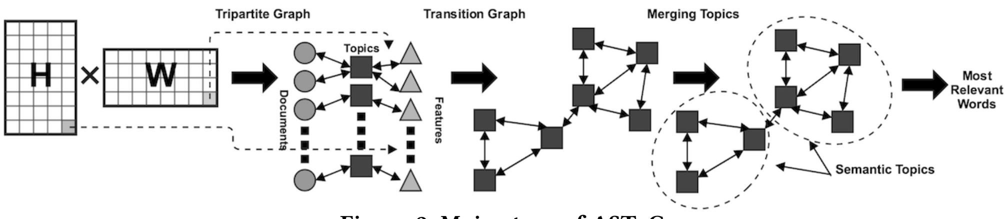


Figure 2: Main steps of ASToC.

4.1 Experimental Setup

4.1.1 Datasets. The primary goal of our solution is to perform effectively topic modeling so that more coherent topics can be extracted. To evaluate topic model coherence, we consider 12 real-world datasets as a reference. Two of them were created by us by collecting comments from Facebook and Uber Apps in Google Play Store. The others were obtained from previous works in the literature. For all, we performed stopword removal (using the standard SMART list) and removed words such as adverbs, using the VADER lexicon dictionary [12], as the vast majority of the important words for identifying topics are nouns and verbs. These procedures improved both, the efficiency and effectiveness of all analyzed strategies. Table 1 provides a brief summary of the reference datasets, reporting the number of features (words), documents, the mean number of words per document (density).

Dataset	#Feat	#Doc	Density
Angrybirds [10]	1,903	1,428	7.135
Dropbox [10]	2,430	1,909	9.501
Evernote [10]	6,307	8,273	11.002
InfoVisVast	6,104	909	86.215
Pinterest [10]	2,174	3,168	4.478
TripAdvisor [10]	3,152	2,816	8.532
Tweets [17]	8,029	12,030	4.450
WhatsApp [10]	1,777	2,956	3.103
20NewsGroup [3]	29,842	15,411	76.408
ACM [23]	16,811	22,384	30.428
Uber	5,517	11,541	7.868
Facebook	5,168	12,297	6.427

Table 1: Dataset characteristics

4.1.2 Evaluation, Algorithms and Procedures. We compare the topic modeling strategies using representative topic quality metrics in the literature [19, 20]. In general, there are three class of topic quality metrics based on three criteria: (a) coherence, (b) mutual information and (c) semantic representation. We also consider three topic lengths (5, 10 and 20 words) under each metric in our evaluation—different lengths may bring different challenges.

Regarding the metrics, *coherence* captures easiness of interpretation by co-occurrence. Words that co-occur frequently in similar contexts in a corpus are easier to correlate since they usually define a more well-defined "concept" or "topic". For coherence, we employ an improved version of regular coherence [20], called TFIDF-Coherence, defined as

$$c_{tf-idf}(t, W_t)=\sum_{w_1,\, w_2\in W_t} log\frac{\sum_{d:w_1,\, w_2\in d} tf-idf(w_1, d)tf-idf(w_2, d)}{\sum_{d:w_1\in d} tf-idf(w_1, d)} \quad (5)$$

where the *tf-idf* is computed with augmented frequency as

$$tf-idf(w, d)=\left(\frac{1}{2}+\frac{f(w, d)}{max_{w'\in d} f(w', d)}\right) log\frac{|D|}{|\{d\in D: w\in d\}|} \quad (6)$$

and $f(w, d)$ is the number of occurrences of a term w in document d. This skews the metric towards topics with high *tf-idf* scores since the numerator of the coherence fraction has quadratic dependence on the *tf-idf* scores and the denominator only linear.

Another class of topic quality metrics is based on the notion of *pairwise point-wise mutual information (PMI)* between the top words in a topic. It captures how much one "gains" in information given the occurrence of the other word, taking dependencies between words into consideration. Following a recent work [19], we compute a *normalized version of PMI* (NPMI), in which, for a given ordered set of top words W_t=$(w_1, ..., w_N)$ in a topic, NPMI is computed as

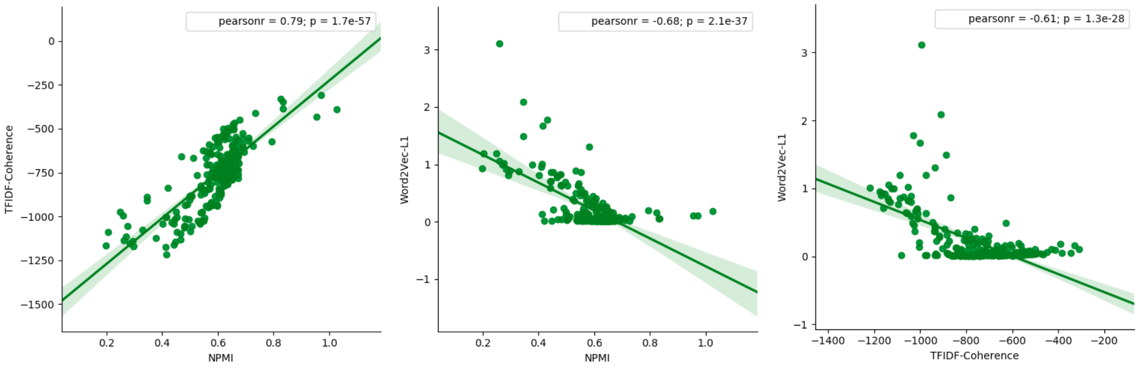
$$NPMI_t=\sum_{i<j}\frac{log\frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-log\, p(w_i,\, w_j)} \quad (7)$$

Finally, the third class of metrics is based on distributed word representations introduced in [19]. The intuition is that, in a well-defined topic, the words should be semantically similar, or at least related, to be easily interpretable by humans. Hence, in a vector space model R^d in which every vocabulary word $w\in W$ has been assigned to a vector v_w, the vectors of top words in a topic should be close to each other. In [19], the authors define topic quality as the average distance between the top words in the topic, $W2V{-}L1=\frac{1}{|W_t|(|W_t|-1)}\sum_{w_1\neq w_2\in W} d(v_{w1}, v_{w2})$.

If $d(w_1, w_2)$ is a distance function in R^d, larger results correspond to worse topics (with words not as localized as in topics with smaller average distances). In [19], they suggest four different distance metrics, with Euclidean (L_1) distance presenting the best results. We also employ the L_1 distance, defined as $d_{L_1}(x, y)= \sum_{i=1}^{d} |x_i{-}y_i|$.

Next, we compare several instantiations of our framework using the settings described in Section 3, as well the best configuration with eight topic model strategies recently proposed, marked in bold in Section 2. For all strategies that use word embeddings, we adopt the Word2Vec representation trained with the GoogleNews [18], with exception of the ETM strategy that was trained with Glove [21], as suggested by the authors. We assess the statistical significance of our results by means of a paired t-test with 95% confidence and Holm-Bonferroni correction to account for multiple tests. We present all results in Figure 4, one graphic for each pair of metric/topic length considered. In the next section, we focus our discussions on the comparison of the results obtained with the several instantiations of our framework. Then we compare the best instantiation (*FV+NMF+ASToC*, as we shall see) with the other eight baseline strategies.

4.1.3 Correlation between evaluation metrics. We provide a brief analysis to verify whether there is any correlation between the three

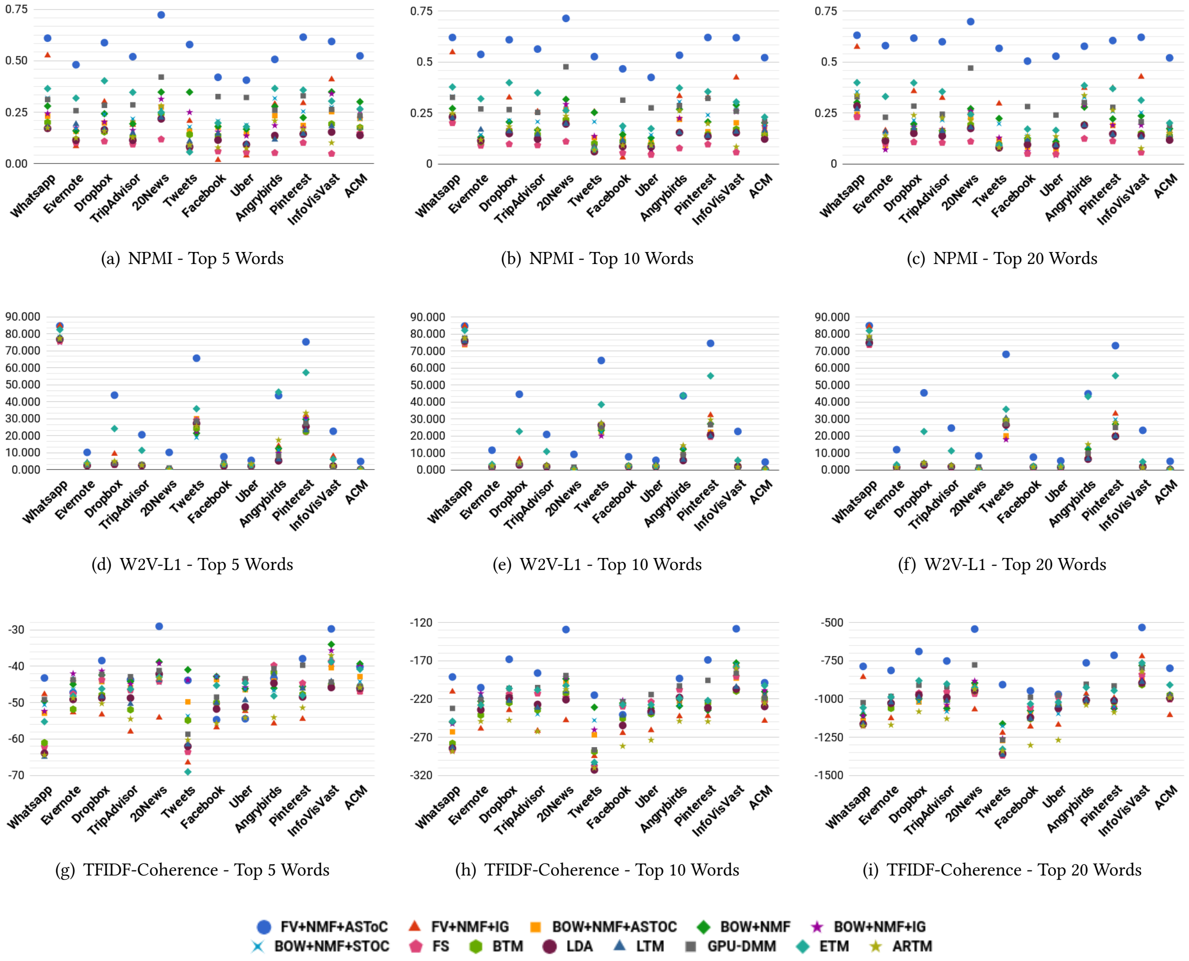


(a) TFIDF-Coherence x NPMI

(b) W2V-L1 x NPMI

(c) W2V-L1 x TFIDF-Coherence

Figure 3: Linear correlation between the evaluation metrics. TFIDF-Coherence x NPMI present a monotonically increasing relationship while W2V-L1 x NPMI and W2V-L1 x TFIDF-Coherence present a monotonically decreasing relationship.



(a) NPMI - Top 5 Words

(b) NPMI - Top 10 Words

(c) NPMI - Top 20 Words

(d) W2V-L1 - Top 5 Words

(e) W2V-L1 - Top 10 Words

(f) W2V-L1 - Top 20 Words

(g) TFIDF-Coherence - Top 5 Words

(h) TFIDF-Coherence - Top 10 Words

(i) TFIDF-Coherence - Top 20 Words

Figure 4: Comparing the results achieved by each strategy considering top 5, 10 and 20 words for NPMI, W2V-L1 and TFIDF-Coherence. The *FV+NMF+ASToC* instantiation always achieves the best results, with few ties with other baseline strategies.

used evaluation metrics used in our evaluation. The Pearson's correlation coefficient is a measure of the strength of the association between metrics. We organize the results as follows: for each dataset c, we trace each s for a metric m in vector $v_{\vec{\mathrm{c},}\,\mathrm{m}}$=$\{s_{c_{t1}}, \cdots, s_{c_{t_k}}\}$. Then, we arrange the 12 datasets in an order $\{c_1, \cdots, c_{12}\}$, to maintain the correctness of the comparison of the metrics, and finally we concatenate the vectors $v_{\vec{\mathrm{c},}\,\mathrm{m}}$ in a single vector $\vec{v_{\mathrm{m}}}$=$\{v_{\vec{c_1,}\,\mathrm{m}}, \cdots, v_{\vec{c_{12},}\,\mathrm{m}}\}$. Figure 3 shows the pairwise correlation among the evaluation metrics. Figure 3(a) shows a strong increase correlation among TFIDF-Coherence and NPMI, with Pearson score (*pearsonr*) equals to 0.79 and a low probability (p) the two metrics have no correlation. On the contrary, Figures 3(b) and 3(c) show a decrease linear correlation with *pearsonr* equals to -0.68 and -0.61, respectively. These results show that, though moderatedly correlated, each metric explores different aspects in the assessment of the quality of topics, justifying the importance of using all of them.

4.2 Experimental Results

4.2.1 Effectiveness of the instantiations of our proposed framework. We turn now our attention to the effectiveness of distinct configurations of our proposed framework described in Section 3. We report NPMI, W2V-L1 and TFIDF-Coherence results of the selected framework variations in the graphics of Figure 4. We discovered 25 topics for all datasets except 20News, ACM and Tweets, where 20, 11 and 6 topics were discovered, respectively. The number of topics for the app datasets was defined based on the choice of topics made in [10]. For the 20News, ACM and Tweets datasets, we chose the number of topics equals to the real number of classes. Considering the statistical significance tests, the *FV+NMF+ASToC* configuration is the top-performer amongst the explored variations. Such configuration achieves the best results in 11 out 12 datasets, considering the NPMI for top 5, 10 and 20 words, tying in the Tweets for top 5 and 10 words with *BOW+NMF* and *BOW+NMF+IG* and top 20 words with *BOW+NMF*. *FV+NMF+ASToC* improves the results of *BOW+ASToC* due to its richer data representation strategy. *FV+NMF+IG* is the worst setting among the evaluated variations of our framework. We argue that this has to do with the fact that *FV+NMF+IG* uses only one of the residual matrices produced by NMF to discover the topics. Using the Fisher Vector representation, NMF alone is not able to exploit all the information factored to infer topics using only a single residual matrix, since the information about words is not well defined as in the BOW representation. *FV+NMF+ASToC*, on the other hand, is capable of mitigating this issue since *ASToC* uses the two residual matrices to discover topics, taking full advantage of all information provided by the Fisher Vector representation. Similar results can be observed for W2V-L1.

We should notice that in the case of the tf-idf coherence metric, we have more ties in comparison to the two other classes of topic quality metrics. We explain this by use of the own tf-idf scores in the input matrix of the latent topic decomposition models. This causes a bias towards high tf-idf words in all models (some of ours and the baselines). On the other hand, our best strategy (*FV+NMF+ASToC*) uses the word embeddings representation and, consequently, the topics are defined based on the richer semantics carried by this representation. Thus, this strategy captures tf-idf coherence information such as the baseline techniques that make direct use of that information.

4.2.2 Effectiveness results Against the Baselines. We compare our framework's best configuration – *FV+NMF+ASToC* – a-gainst eight state-of-the-art topic modeling strategies considering the twelve reference datasets. Once again, in Figure 4 we contrast the results of our proposed framework and the reference strategies, considering NPMI, W2V-L1 and TFIDF-Coherence metrics. We consider the same values of k reported in the previous experiment. As we can see, our strategy achieves statistically significant gains in terms of the quality of the discovered topics in all 12 datasets, considering the NPMI score. Most baselines cannot get even close, with the best baseline (GPU-DMM) being worse than *FV+NMF+ASToC* by more than 30% when it obtains its best performance (in 20news). Considering the TFIDF-Coherence scores, our strategy does not achieve the best results just for Facebook and Uber for top 5 and 10 words, but it is very close to those. For all remaining datasets, we achieved the best results or tie with the best baselines, with gains of more than 70% against the runnerup (GPU-DMM) in some datasets. Similar results can be observed for W2V-L1 metric (gains achieve up to 80%). Considering the comparison with GPU-DMM and ETM, we can notice that embeddings can indeed enrich the model representation, but our results reinforce that directly exploiting them in the data representation is better than using them as auxiliary information.

Method	Metric			Σ
	NPMI	W2V-L1	TFIDF	(Sum)
FV+NMF+ASToC	**36**	**31**	**32**	**99**
FV+NMF+IG	2	8	5	15
BOW+NMF+ASTOC	0	6	14	20
BOW+NMF	4	6	12	22
BOW+NMF+IG	2	8	14	24
BOW+NMF+STOC	1	7	11	19
FS	0	6	11	17
BTM	1	7	7	15
LDA	0	6	8	14
LTM	0	6	9	15
GPU-DMM	**2**	**9**	**19**	**30**
ETM	1	15	10	26
ARTM	0	11	4	15

Table 2: Number of times each strategy was the top performer. $FV{-}NMF{-}ASToC$ **is the choice in the vast majority of cases.**

We provide a comparative table with all experimental results[4], In the Table, statistical significance tests ensure that the best results, marked in ▲, are superior to others. The statistically equivalent results are marked in •. We summarize our findings regarding the behavior of all analyzed strategies in the 12 datasets in Table 2. It counts the number of times each strategy figured out as a top performer[5]. As we can see, our proposal is in great advantage over the other explored baselines, being the strategy of choice in the vast majority of cases. Overall, considering a universe of 108 experimental results (combination of 3 evaluation metrics with the 3 topic lengths over 12 datasets) we obtained the best results (99 best performances), with the GPU-DMM coming far away, with just 30 top performances.

4.3 Quantitative Analysis

One important question remains to be answered: to what extent each component of the *FV+NMF+ASToC* configuration contributes

[4] *https://drive.google.com/file/d/10CQ7e-TVkxIo1fSnc9CHweP4PRvkDRXg*

[5] If two strategies are statistically tied as top performers in the same dataset, both will be counted.

to the best results? In order to answer this question and gather insights about the properties of the proposed framework, we provide a quantitative analysis regarding the effects of its main components on topic modeling effectiveness (measured by the NPMI score). We not only account for the effects of each component of the scores, but also for the variability due to the intrinsic noise observed in textual data.

We start our analysis by quantifying the effects of various factors (parameters of interest) that might affect the performance of the system under study, while also determining whether the observed variations are due to significant effects or simply to random variations (e.g., measurement errors, inherent variability of the process being analyzed [13]). To this end, we adopt a $2^k r$ factorial design, since it allows us to quantify the effect of k factors (and their interactions), under two possible levels (lower and upper), on the performance of the system under study (measured by a given response variable). A $2^k r$ design consists of 2^k experiments defined by all combinations of factor levels. Each experiment is repeated r times to account for the inherent variability in the data. Such design is appropriate to provide some initial insights into the relative importance of different factors (and factor interactions) on the response variable. This importance is expressed as the fraction of the total variation observed in the measurements that can be explained by the changes in the levels of each factor (and interaction): the more important factors explain a larger fraction of the total variation. The fraction of variation that cannot be explained is credited to the experimental errors.

We here apply a $2^k r$ design to analyze the impact of the main characteristics of our proposed strategy. The ultimate goal is to better understand the main factors that contribute to the generation of more cohesive topics. As previously mentioned, the response variable in our design is the average NPMI score, considering all explored datasets. Our factorial design consists of r=12 replications (one per dataset). We evaluate the two most significant factors (k=2) that are capable of affecting the results of the proposed approach. The first one is the use of ASToC to group semantically similar topics. The second factor is the use of the Fisher Vector representation (which relies on the word selection of topics using IG).

Having defined the k=2 factors of our design, the next step is to evaluate the four (2^k) variations of our general framework. For the ASToC factor, the upper level corresponds to the use of ASToC in the experiments, while the lower one corresponding to its absence. Likewise, the use of the Fisher Vector representation is the upper level of the second evaluated factor, with the lower level corresponding to its replacement by the traditional BOW representation. All four configurations considered in the experimental evaluation are as follows: (i) absence – absence → **BoW + NMF**; (ii) presence – absence → **FV + NMF + IG**; (iii) absence – presence → **BoW + NMF + ASToC**; (iv) presence – presence → **FV + NMF + ASToC**.

The results regarding the factorial design can be found in Table 3. As we can see, a significant amount (31%) of the observed variations in the obtained scores can be attributed to the use of ASToC. This provides strong evidence in favor of the potential of using ASToC as a mechanism to improve topic models by grouping semantically similar topics in a principled way. The interaction between ASToC and Fisher Vector (FV) factors also explains a large proportion (22%) of the observed variations, suggesting that the data representation directly affects the ASToC strategy for topic merging. In fact, this interaction is more relevant than the data representation alone, since the Fisher Vector factor explains only 12% of the observed variations. Finally, a large amount of the observed variations (35%) was credited to experimental errors, remaining as inexplicable variations. This has to do with the large NPMI scores variations observed for each evaluated sample (especially in the smaller datasets).

Fisher Vector	ASToC	Interaction	Inexplicable
12%	31%	22%	35%

Table 3: Factorial design results.

In order to further understand the inexplicable variability in our experiments, we perform a variability analysis using measures that take into account the outliers and skewed values of the obtained scores. Particularly, we consider the median and interquartile range (IQR), which can be seen as alternative measures to the previously presented mean and standard deviation. IQR measures the difference between the third and the first quartiles (a.k.a., the 75th and 25th percentiles, respectively). Similarly to the mean and standard deviation metrics, the median and IQR metrics measure the central tendency and spread (variability) of a set of values. However, they have the advantage of being robust to outliers and non-normal data. More specifically, the two main advantages of median and IQR are: (i) estimating outliers by looking at values at least $1.5 \times IQR$ below the first quartile or above the third quartile; (ii) assessing data skewness by comparing the median to the quartile values. We analyze the obtained NPMI scores on the top 20 words considering the median, IQR, and number of outliers regarding the NPMI scores obtained for the evaluated samples. Table 4 summarizes the results considering for all datasets and the two best performance topic modeling approaches.

Measures	**FV+NMF+ASToC**	**GPU-DMM**
Median	2.422 ▲	0.852
IQR	0.821 ●	1.047 ●
Outliers	1.333 ●	1.454 ●

Table 4: Skewness analysis regarding the NPMI scores.

The same conclusions about the effectiveness of the proposed framework using the mean of NPMI scores remain valid when considering the median: the median value of the NPMI scores obtained for the proposed framework is significantly higher than the median values obtained for the second best performance method. Thus, this analysis complements our previous one (based on the mean of the scores), providing additional evidence in favor of the benefits of combining both, Fisher Vector and ASToC. Despite the fact that the difference between the third and first quartiles (IQR) of the proposed method is higher than the IQR of the other, there is no statistical significance on IQR values in GPU-DMM approach. Furthermore, all IQR values are reasonably high when compared to the corresponding median values. We conjecture that this is due to the fact that it is not trivial to obtain homogeneous NPMI distributions among distinct topic models, since the number of discriminative patterns related to each topic may vary, contributing to the higher variability of the observed results. Finally, our analysis shows that there is a negligible number of outliers in the scores for both topic modeling approaches. This provides evidence that the reported results are not biased by outliers.

4.4 Application: Document Classification

Our proposed method is able to generate more cohesive topics and potentially better document representations, which can help in

Data Repr.	ACM			20NewsGroup		
	MacroF1	MicroF1	Rel. Feat.	MacroF1	MicroF1	Rel. Feat.
TFIDF	60.90 (0.96)	74.29 (0.56)	strict, condition, devs, starter, tomasz	86.96 (0.77)	87.33 (0.71)	edu, know, article, thanks, writes
TFIDF + ASToC	**62.65 (1.05)** ●	**76.20 (0.53)** ●	l_2, l_7, l_5,l_1, l_{10}	**88.08 (0.61)** ▲	**88.50 (0.56)** ▲	l_3, l_5, l_9, l_1, l_{19}
TFIDF + GPU-DMM	**63.14 (1.02)** ●	**76.87 (0.40)** ●	l_2, l_7, l_5, l_1, l_{10}	86.98 (0.51)	87.38 (0.57)	$l18$, l_{19}, l_6, l_{14}, l_{10}

Table 5: Efficacy results of ASToC and GPU-DMM in task classification using the BERT classifier.

tasks such as classification and clustering. Due to lack of space, we analyze the suitability of our model in the classification task, leaving the analysis of other applications for future work. We consider the ACM and 20Newsgroup datasets as they are the only evaluated datasets commonly adopted in this task. We compare our proposed method with the best baseline, GPU-DMM. For both FV + NMF + ASToC and GPU-DMM, we set the parameter k equal to the number of classes of each dataset. We used 5-fold cross-validation and the BERT classifier [4]. We adopt BERT, a novel Random Forest based state-of-the-art classifier (as good or better than SVM) that improves classification generalization by exploiting out-of-bags and extreme randomization [4]. BERT can potentially better benefit from the latent attributes attributes than SVMs, which are very resilient to new information given their capability of effectivelly exploiting high dimensional sparse spaces. We set the BERT trees to maximum depth and the other parameters were tuned via 5-fold nested cross-validation within the training set. For our FV + NMF + ASToC model, we represent each document using the latent vector information that represents the respective document combined with the TFIDF representation of the document words. The same data representation was performed for the GPU-DMM. We also consider only the TFIDF representation as the baseline.

We assess the statistical significance of our results by means of a paired t-test with 95% confidence and Holm correction to account for multiple tests. This test assures that the best results, marked with ▲, are statistically superior to others. Statistical ties are represented with ●. Table 5 presents MacroF1 and MicroF1 results, showing that our method is able to improve the representation of the documents when compared to the GPU-DMM and the original TFIDF representation. We can also observe in the column Rel. Feat. that the five most relevant features for all representations that use information about the topics are latent attributes l_i, where i corresponds to the latent topic.

5 CONCLUSIONS AND FUTURE WORK

We proposed a new topic modeling framework capable of finding high-quality topics on textual data. Our experimental results provide strong evidence in favor of adopting a richer data representation in conjunction topic merging based on semantic similarity. One specific configuration of our proposed framework, *FV+NMF+ASToC*, was able to achieve consistently better results than recently proposed state-of-the-art topic modeling techniques in all explored datasets. In fact, from all 108 evaluated situations, we won or tied in basically in 99 of them (against 30 of the runner-up), with gains, for instance, of more than 30% in coherence, 50% in mutual information and almost 80% in semantic similarity. The analysis of the main innovations of our framework (namely the Fisher Vector representation and ASToC) highlights the importance of both for effective topic modeling, as well as the intrinsic variability of the task. However, it also shows where there is room for further improvements. As future work, we plan to invest in removing noisy and irrelevant topics. We also plan to assess the impact of using our high-quality topics on downstream algorithms, such as sentiment detectors and recommendation systems. Two interesting research directions are to assess the impact of using other matrix factorization techniques such as the SVD decomposition and to test our model in other applications such as clustering.

6 ACKNOWLEDGMENTS

This work is partially supported by CAPES, CNPq, Finep, Fapemig, Mundiale, Astrein, projects InWeb and MASWeb.

REFERENCES

[1] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL'14*.

[2] P. V. Bicalho, T. de Oliveira Cunha, F. H. J. Mourão, G. L. Pappa, and W. M. Jr. Generating cohesive semantic topics from latent factors. In *BRACIS*, 2014.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 2003.

[4] R. Campos, S. Canuto, T. Salles, C. C. de Sá, and M. A. Gonçalves. Stacking bagged and boosted forests for effective automated classification. In *SIGIR*, 2017.

[5] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *SDAIR'94*, 1994.

[6] Z. Chen and B. Liu. Topic modeling using topics from many domains, lifelong learning and big data. In *ICML'14*, 2014.

[7] X. Cheng, X. Yan, Y. Lan, and J. Guo. Btm: Topic modeling over short texts. *IEEE TKDE '14*, 2014.

[8] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE TVCG*, 2013.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIST*, 1990.

[10] E. Guzman and W. Maalej. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering*, 2014.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, 1999.

[12] C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM'14*, 2014.

[13] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. 1991.

[14] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[15] G. Lev, B. Klein, and L. Wolf. *NLDB'15*, chapter In Defense of Word Embedding for Generic Text Representation. Springer International Publishing, 2015.

[16] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM TOIS*, 2017.

[17] Q. Li, S. Shah, X. Liu, A. Nourbakhsh, and R. Fang. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *CIKM'16*.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. 2013.

[19] S. I. Nikolenko. Topic quality metrics based on distributed word representations. In *SIGIR'16*, 2016.

[20] S. I. Nikolenko, S. Koltcov, and O. Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 2017.

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[22] J. Qiang, P. Chen, T. Wang, and X. Wu. Topic modeling over short texts by incorporating word embeddings. In *PAKDD*. Springer, 2017.

[23] F. Viegas, M. Gonçalves, W. Martins, and L. Rocha. Parallel lazy semi-naive bayes strategies for effective and efficient document classification. In *CIKM*, 2015.

[24] K. Vorontsov and A. Potapenko. Additive regularization of topic models. *Mach. Learn.*, 101(1-3):303–323, 2015.

[25] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 2004.