PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 115 OCTOBER 2020 31–50

Inferring Highly-dense Representations for Clustering Broadcast Media Content

Esaú Villatoro-Tello,[a b] Shantipriya Parida,[b] Petr Motlicek,[b] Ondřej Bojar[c]

[a] Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
[b] Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland.
[c] Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25 118 00 Praha 1, Czech Republic

Abstract

We propose to employ a low-resolution representation for accurately categorizing spoken documents. Our proposed approach guarantees document clusters using a highly dense representation. Performed experiments, using a dataset from a German TV channel, demonstrate that using low-resolution concepts for representing the broadcast media content allows obtaining a relative improvement of 70.4% in terms of the Silhouette coefficient compared to deep neural architectures.

1. Introduction

Current broadcast platforms utilize the Internet as a cross-promotion source, thus, their produced materials tend to be very short and thematically diverse. Besides, modern Web technologies allow the rapid distribution of these informative content through several platforms. As a result, the broadcast media content monitoring represents a challenging scenario for current Natural Language Understanding (NLU) approaches to efficiently exploit this type of data due to a lack of structuring and reliable information associated with these contents (Morchid and Linarès, 2013; Doulaty et al., 2016; Staykovski et al., 2019). Furthermore, if we consider that documents are very short and that they come from a very narrow domain, the task of clustering becomes harder.

Traditionally, the Bag-of-Words (BoW) has been the most widely used text representation technique for solving many text-related tasks, including document cluster-

Corresponding author: shantipriya.parida@idiap.ch
Cite as: Esaú Villatoro-Tello, Shantipriya Parida, Petr Motlicek, Ondřej Bojar. Inferring Highly-dense Representations for Clustering Broadcast Media Content. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 31–50. doi: 10.14712/00326585.004.

ing, due to its simplicity and efficiency (Ribeiro-Neto and Baeza-Yates, 1999). However, the BoW has two major drawbacks: *i*) document representation is generated in a very high-dimensional space, *ii*) it is not feasible to determine the semantic similarity between words. As widely known, previous problems increase when documents are short texts (Li et al., 2016). It becomes more difficult to statistically evaluate the relevance of words given that most of the words have low-frequency occurrences, the BoW representation from short-texts results in a higher sparse vector, and the distance between similar documents is not very different than the distance between more dissimilar documents.

To overcome some of the BoW deficiencies, semantic analysis (SA) techniques attempt to interpret the meaning of the words and text fragments by calculating their relationship with a set of predefined concepts or topics (Li et al., 2011). Examples of SA techniques are LDA (Blei et al., 2003), LSA (Deerwester et al., 1990), and word embeddings (Le and Mikolov, 2014; Bojanowski et al., 2017; Devlin et al., 2019). Accordingly, these strategies learn word or document representations based on the combination of the underlying semantics in a dataset. Similarly, more recent approaches, with the help of word embeddings, learn text representations using deep neural network architectures for document classification (De Boom et al., 2016; Adhikari et al., 2019; Ostendorff et al., 2019; Sheri et al., 2019). However, most of these approaches focus either on solving supervised classification tasks or clustering formal-written short documents.

In this paper, we propose an efficient technological solution for the unsupervised categorization of broadcast media content, i.e., spoken documents. Our proposed approach generates document clusters using a highly dense representation, referred to as low-resolution concepts. We first identify the fundamental semantic elements (i.e., concepts) in the document collection, then, these are used to build the low-resolution representation, which is later used in an unsupervised categorization process. One major advantage of our proposed approach is it's easy to interpret, explicit, and profound representation, allowing the end-users understanding of document vectors and their differences.

The main contributions of this paper are summarized as follows: *i*) To the best of our knowledge, this is the first attempt to explore the feasibility and effectiveness of the low-resolution bag-of-concepts in solving one particular unsupervised task, broadcast media content categorization; *ii*) We conducted our experiments on a real-life dataset of German spoken documents, achieving good performance in terms of three internal evaluation metrics, allowing our method to be considered for practical deployment; *iii*) We evaluate the performance of our proposed method in three well-known datasets (formal written documents).

The remainder of the paper is organized as follows: a brief description of the related work is given in Section 2, in Section 3 we describe the proposed methodology, Section 4 we provide some details regarding the employed dataset. Experimental results and analyses are presented in Sections 5 and 6. Finally, in Section 7 we draw our main conclusions and future work directions.

2. Related Work

Our work is mainly related to topic modeling or topic discovery. As known, topic discovery aims to use statistical information of word occurrences to obtain the underlying semantics contained in a document set. The most popular textual topic modelling are based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Bayesian methods represented by latent semantic analysis (LSA) (Deerwester et al., 1990), Hierarchical Dirichlet Process (HDP) (Teh et al., 2005).

During recent years, models based on deep neural networks have emerged as a viable alternative for topic discovery. For example, the replicated softmax model (RSM), based on Restricted Boltzmann Machines (Hinton and Salakhutdinov, 2009), which is capable to estimate the probability of observing a new word in a document given previously observed words, thus RSM can learn efficient document representations. More recently, Variational Autoencoders (VAEs) have been successfully adapted for text topic modeling. The Neural Variational Document Model (NVDM) (Miao et al., 2016) for text modeling is an extension of a standard VAE, with an encoder that learns Gaussian distribution and a softmax decoder capable of reconstructing documents in a semantic word embedding space. In (Silveira et al., 2018) authors propose a VAE-based on Gumbel-Softmax (GSDTM) and Logistic-normal Mixture (LMDTM) for text topic modelling. In (Wang et al., 2020) authors propose a neural topic modeling approach, called Bidirectional Adversarial Topic (BAT) model, which builds a two-way projection between the document-topic distribution and the document-word distribution. Although these recent approaches have demonstrated great improvement in text clustering tasks using the topic information, they all have one major disadvantage, they require great amounts of data to infer accurate semantic representations, plus the lack of interpretability.

Despite the extensive exploration of this research field, scarce work has been done to evaluate the impact of these technologies in speech-documents, i.e., textual transcriptions obtained from speech. Contrary to formal documents, textual transcript represents a more challenging scenario as they represent very short documents, containing several speech phenomena such as hesitation, fillers, repetition, etc. Accordingly, in this paper, we evaluate the impact of several clustering strategies for broadcast media categorization. Our proposed approach generates document clusters using highly dense representation, which are easy to interpret by a human judge. The recent relevant work to ours is proposed by (Kim et al., 2017), which proposes a bag-of-concepts approach to generate alternative document representations to overcome the lack of interpretability of word2vec and doc2vec methodologies. However, contrary to this particular work, our method is particularly suited for very short spoken documents (transcripts), and we use highly dense representations, i.e., a very small

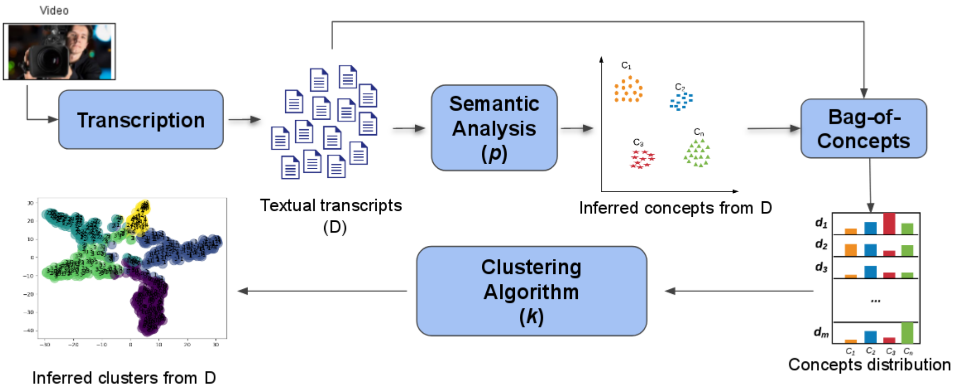

Figure 1: General framework to categorize spoken-documents using low-resolution concepts.

set of features is used to represent the concepts contained in the dataset. We evaluate our proposed method in a real-life dataset extracted to form a German tv channel and we also evaluate our method's performance in three benchmark corpora.

3. Proposed Method

Inspired by the work of (Kim et al., 2017; López-Monroy et al., 2018), we propose using a highly dense representation, denominated low-resolution concepts, for solving the task of clustering short transcript-texts, i.e., broadcast media documents. The intuition behind this approach is that highly abstract semantic elements (concepts) are good discriminators for clustering very short transcript texts that come from a narrow domain. The proposed methodology is depicted in Figure 1. Generally speaking, we first identify the underlying concepts contained in the dataset. For this, we can employ any semantic analysis (SA) approach for learning words representation; thus, learned representation allows us to generate sets of semantically associated words. After obtaining the main concepts, documents are represented by a condensed vector, which counts for the occurrences of the concepts, i.e., a concept distribution vector. Finally, the build texts representation serves as the input to a clustering process, in this case, the K-means algorithm.

More formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ denote the set of short transcript texts, and let $\mathcal{V} = \{w_1, w_2, \ldots, w_m\}$ represent the vocabulary of the document collection $\mathcal{D}$. As first step, we aim at inferring the underlying set of concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_p\}$ contained in $\mathcal{D}$, where every $c_l \in \mathcal{C}$ is a set formed by semantically related words. Notice that in order to obtain the concepts $\mathcal{C}$ we can apply any SA technique for learning the vector representation $\mathbf{v}_i$ of each word $w_i \in \mathcal{V}$, for example LDA, LSA, or word embeddings. Next, for obtaining the document $\mathbf{d_j}$ representation, we account for the occurrence of each c_l within d_i, in other words, the document vector $\mathbf{d_j}$ is a vector that contains concepts distribution. Finally, the generated document-concepts matrix $\mathbf{M}_{\mathcal{D}\times\mathcal{C}}$ serves as the input to a clustering process aiming at finding the more suitable documents groups according to the concept-based representation. Henceforth, we will refer to the document-concepts matrix as the Bag-of-Concepts (BoC) representation.

The proposed method has two main parameters, the resolution parameter (p) and the group parameter (k). The former, p, represents the number of concepts that will be generated from the SA step. The lower the number of concepts, the more abstract the resolution. The second parameter, k, indicates the number of categories to be generated from the clustering process. Given the nature of the dataset, i.e., very short texts from a narrow domain, we hypothesize that the clustering algorithm will be able to find groups of documents that share the same amount of information about the same sub-set of concepts, resulting in a more coherent categorization of the documents. Thus, using low-resolution concepts will generate groups of documents referring to the same general topics, while using higher resolution values will result in a more fine-grained topic categorization of the documents.

4. Dataset Description

The dataset used in our paper is from n-tv[1], a German free-to-air television news channel. There are mainly two different sets of files in the proprietary data. One part of the dataset is represented by the speech segments (audio data) with an average duration of 1.5 minutes where each recording has multiple speakers recorded in a relatively noisy environment. The other part of the dataset is the textual transcripts (German) associated with the speech segments. Each of the transcript files represents an article (short text documents), which usually are spread across different topics. See for example a small fragment of an article shown in Table 1. This example, when given to experts, is categorized as 'politics' and as an 'economy' article, which is somehow correct given that both topics are present in the article. This occurs repeatedly across articles due to the interviewed people often mix topics when spontaneously speaking, making the categorization task even more challenging.

For our experiments, the employed dataset comprises a total of 697 articles. Table 2 shows some statistics from the employed dataset; before applying any pre-processing operation and after pre-processing. As pre-processing operations, we removed stopwords, numbers, special symbols, all the words are converted to lower-case. [2]We compute the average number of tokens, vocabulary, and lexical richness (LR) in the dataset. A couple of main observations can be done at this point. On the one hand,

[1] https://www.n-tv.de/

[2] We did not make any special processing for German compounds words.

Original German fragment
Arbeitsminister Hubertus Heil **kämpft** für **befristete Teilzeit**. Also dafür dass man nicht nur von Voll-zur Teilzeit sondern eben auch wieder zurück wechseln kann ...der **Arbeitgeber** darf den Antrag auf Teilzeit auch nicht einfach so ausschlagen außer es gibt betriebliche Gründe... bei **Unternehmen** mit mehr als 200 **Mitarbeitern** habe alle ein Recht auf befristete Teilzeit...zudem kann der Arbeitgeber den Antrag auf befristete Arbeitszeit ablehnen wenn diese ein Jahr unter- oder fünf Jahre überschreitet.
Closest English translation
Minister of Labor Hubertus Heil is **fighting** for **part-time work**. So that you can not only switch from full-time to part-time but also back again ... the **employer** may not simply refuse the application for part-time unless there are operational reasons ... in **companies** with more than 200 **employees**, everyone has a right to temporary part-time work ... the employer can also reject the application for limited working time if it exceeds one year less than or five years.

Table 1: Extracted fragment from the n-tv dataset. Letters in **bold** represent keywords associated with *politic* and *economic* topics.

we notice that individual texts are very short, on average 63.02 tokens with an average vocabulary of 47.86 words, resulting in a very high LR (0.785). This suggests that very few words are repeated within one article, very few redundancies, which represents a challenge for frequency-based methods. On the other hand, globally speaking, the complete dataset has an LR=0.272, which indicates, to some extent, that the information across texts is highly overlapped (narrow domains).

4.1. Benchmark datasets

To validate our proposal, we also evaluate our method in the following three benchmark datasets:

- **AG's news corpus.** We used the as employed in (Zhang et al., 2015). It contains categorized news articles (4 classes) from more than 2000 news sources. In total, this dataset contains 120000 documents in the train partition and 7600 in the test partition.
- **Reuters.** These documents appeared on the Reuters newswire in 1987 and were manually classified by personnel from Reuters Ltd. Particularly, we used for our experiments the R8 partition as provided in (Cardoso-Cachopo, 2007), i.e., 5845 documents for training, and 2189 for testing divided into eight categories.
- **10KGNAD.** This dataset, based on the One Million Posts Corpus (Schabus et al., 2017), is composed of 10273 German news articles collected from an Australian online newspaper. News is categorized into 9 different topics. The train partition contains 9245 documents, while the test partition contains 1028 documents.

	W/O Pre-processing Average (σ)	Total
Tokens	234.68 (± 124.45)	163,572
Vocabulary	161.79 (± 51.92)	22078
LR	0.717 (± 0.073)	0.134
	W/ Pre-processing Average (σ)	Total
Tokens	63.02 (± 31.52)	43,928
Vocabulary	47.86 (± 16.30)	11,948
LR	0.785 (± 0.092)	0.272

Table 2: Statistics of the n-tv dataset.

5. Experimental framework

This section describes the experimental setup. First, we describe the employed methods for learning word representations. Then, we briefly explain the evaluation metrics; and finally, we describe the approaches used for comparison purposes (baselines). For all the performed experiments we ran the k-means algorithm[3] for a range of $k = 2 \dots 15$.

5.1. Obtaining word vectors

One crucial step of our approach is learning word representations, i.e., the semantic analysis process shown in Figure 1. For this, an important parameter is the resolution value (p), which indicates the number of concepts that will be employed for building the document-concepts matrix (BoC). Accordingly, we evaluate four different methods for inferring the set $\mathcal{C}$ ($|\mathcal{C}| = p$):

- **FastText:** Concepts are inferred from applying a clustering process over $\mathcal{V}$, using as word representation pre-trained word embeddings. We used word embeddings trained with FastText[4] (Bojanowski et al., 2017) on 2 million German Wikipedia articles. This configuration is referred as: **BoC(FstTxt)**.
- **BERT:** Similar to the previous configuration but, here we use BERT (Devlin et al., 2019), a very recent approach for getting contextualized textual representations. Thus, we feed every word in $\mathcal{V}$ to BERT and preserve the encode produced by

[3] As implemented in the scikit-learn library: `https://scikit-learn.org/stable/modules/clustering.html`

[4] `https://www.spinningbytes.com/resources/wordembeddings/`

the last hidden layer (768 units) as the word vector. Performed experiments were done using the pre-trained `bert-base-german-cased` model[5]. We refer to this configuration as **BoC(BERT)**.

- **LDA:** Latent Dirichlet Allocation (Blei et al., 2003) assumes that documents are probability distributions over latent concepts, and concepts are probability distributions over words. Thus, LDA backtracks from the document level to identify concepts that are likely to have generated the dataset. We used the Mallet's LDA implementation from Gensim[6]. After obtaining the concepts, we compute the document-concepts distribution over each $\mathrm{d_j}$ for generating the $\mathbf{d_j}$ representation. We refer to this experiment as **BoC(LDA)**.
- **LSA:** Latent Semantic Analysis (Deerwester et al., 1990) is a purely statistical technique that applies singular value decomposition (SVD) to the term-document matrix to identify the 'latent semantic concepts'. We employed the SVD (singular value decomposition) algorithm as implemented in sklearn[7]. Then, document-concepts representation $\mathbf{d_j}$ is obtained similarly to the LDA approach. We refer to this approach as **BoC(LSA)**.

5.2. Comparisons

We compare the proposed methodology against four different approaches:

- **BoW(*tf-idf*):** Short texts are represented using a traditional Bag-of-Words (BoW) considering a *tf-idf* weighing scheme. The top 10,000 most frequent terms are employed for generating the BoW representation. Thus, once we have the document's representation, we applied the traditional k-means algorithm.

  Avg-Emb: Every short text is represented using the average of the word embeddings which are respectively weighted with their *tf-idf* score. This strategy has been considered in previous research as a common baseline (Huang et al., 2012; Lai et al., 2015; Xu et al., 2015). We used the FastText embeddings for this experiment. Similarly to the BoW baseline, once the representation is generated, we applied the k-means algorithm to perform the clustering process.

  BERT: For this, every text is feed through BERT. As the $\mathbf{d_j}$ representation we use the values of the last hidden layer (768 units). We limit the input length to 510 tokens. After generating the BERT encoding of every document, we applied the k-means algorithm.

  CNNs: Contrary to the previous baselines, this is a specific convolutional neural network designed for clustering short texts[8]. The main idea of this method is to

[5] `https://huggingface.co/transformers/pretrained_models.html`

[6] `https://radimrehurek.com/gensim/models/wrappers/ldamallet.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`

[8] As implemented in `https://github.com/zqhZY/short_text_cnn_cluster`

learn deep features representations without using any external knowledge (Xu et al., 2015).

5.3. Evaluation metrics

For validating the clustering performance we employed three internal methods (Rendón et al., 2011), namely Silhouette (s) score (Rousseeuw, 1987), Calinski-Harabasz (CH) (Caliński and Harabasz, 1974), and Davies-Bouldin (DB) (Davies and Bouldin, 1979) index. Generally speaking, these metrics propose different strategies for combining the concepts of cohesion and separation for each point in the formed clusters. The cohesion value measures how closely the points in a cluster are related among them, and the separation value indicates how well a cluster is distinguished from other clusters.

Silhouette (s) score (Rousseeuw, 1987): this metric combines the concepts of cohesion and separation for each point in the formed clusters. The cohesion value measures how closely the points in a cluster are related among them, and the separation value indicates how well a cluster is distinguished from other clusters. Thus, the s score for a point i is computed as shown in expression 1.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where $a(i)$ is the cohesion score between point i and the rest of the points belonging to the same cluster; and $b(i)$ is the separation score, which represents the minimum average distance between point i and all the other points in any other cluster, of which i is not a member. At the end, the silhouette score of the clustering process is given by the mean $s(i)$ over all points. For this particular metric possible values range between -1 and 1, where a positive result indicates a better quality in the clustering.

Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974): given a dataset $\mathcal{D}$ of size n, divided into k clusters, the CH index is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion. The CH index is computed as shown in expression 2.

$$CH = \frac{SS_B}{SS_W} \times \frac{n-1}{n-k} \tag{2}$$

where SS_W is the overall within-cluster variance, and SS_B is the overall between-cluster variance. The SS_W term represents the sum of the within the sum of squares distances of each point in the cluster from that cluster's centroid, and it will decrease as the number of clusters goes up. On the other hand, the SS_B measures the variance of all the cluster centroids from the dataset's centroid. Hence, a big SS_B value means that all centroids from all clusters are spread out, and consequently not too close to each other. Therefore, the biggest the CH index, the better the clustering output.

Davies-Bouldin (DB) index (Davies and Bouldin, 1979): this index aims to identify sets of clusters that are compact and well separated. The DB index is defined in expression 3.

$$DB = \frac{1}{k} \sum_{i,j=1}^{k} \max_{i \neq j} \left(\frac{d(i, c_i) + d(j, c_j)}{d(c_i, c_j)} \right) \tag{3}$$

where k denotes the number of formed clusters, i and j are cluster labels, then $d(i, c_i)$ is the average distance between each point of cluster i and the centroid of that cluster c_i, this is also know as cluster diameter. Likewise, $d(c_i, c_j)$ is the distance between centroids of cluster i and j respectively. Thus, the smaller the value of the DB index, the better the clustering solution.

Finally, it is worth mentioning that for the experiments performed in the AG's news, Reuters, and 10KGNAD datasets, we evaluate all the possible configurations and baselines on the test partition. Given that these datasets are labeled, we report the obtained results in terms of accuracy (ACC).

6. Results

First, we determine the impact of the resolution parameter (p) in the clustering task. Then, we compare the proposed method using the best value of p against methods described in section 5.2.

6.1. Impact of the resolution

In Figure 2 and Figure 3 we visually show the performance of the considered concepts-inferring approaches in the clustering task, i.e., BoC(FstTxt), BoC(BERT), Boc(LDA), and BoC(LSA). Each map depicts the performance of the different methods under several resolution values $p = 5, 10, 20, 50, 100, 500, 1000$ (y-axis), and several required clusters $k = 2, \dots, 15$ (x-axis). In all cases, the darker the red color in the heat-map the better the performance, conversely, the darker the blue color the worst the performance, and if the cells tend to be white, it means an average performance. Each row in Figure 2 and Figure 3 represents the obtained performance under a different evaluation metric, s score, CH and DB index respectively. As mentioned, the lower the value of the DB index, the better the output of the clustering process. Thus, to provide the generated maps under the third row the same interpretation, we subtract the maximum obtained value under the DB metric to each of the original results.

From these experiments we observe the following: (*1*) Using low-resolution values ($p = 5, 10$) allows us to obtain better performance, showing a consistent behavior

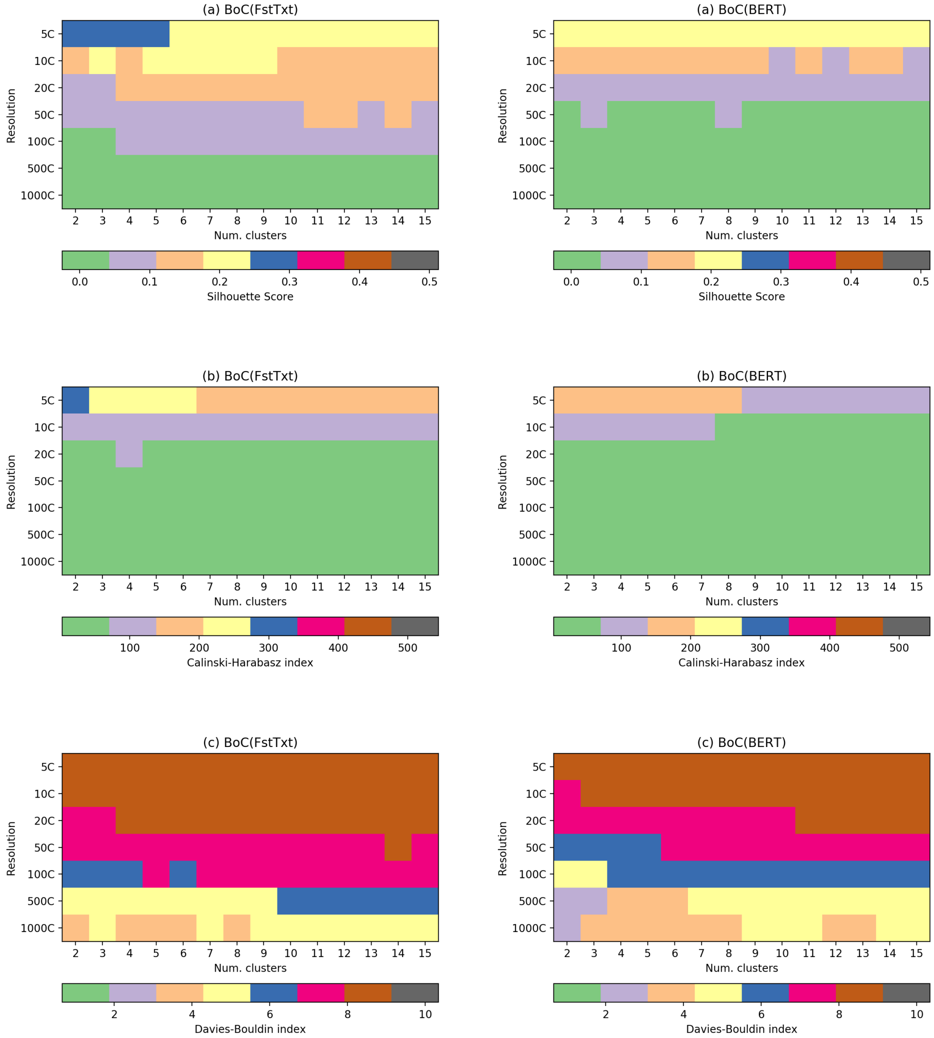

Figure 2: Heatmaps showing the impact of the resolution parameter (p) in the clustering task. First row depicts results in terms of the s score, second row shows the CH index, and third row represents the DB index. Graphs in the same column were generated using the same approach for inferring word representations, specifically, here we are comparing FastText and BERT approaches.

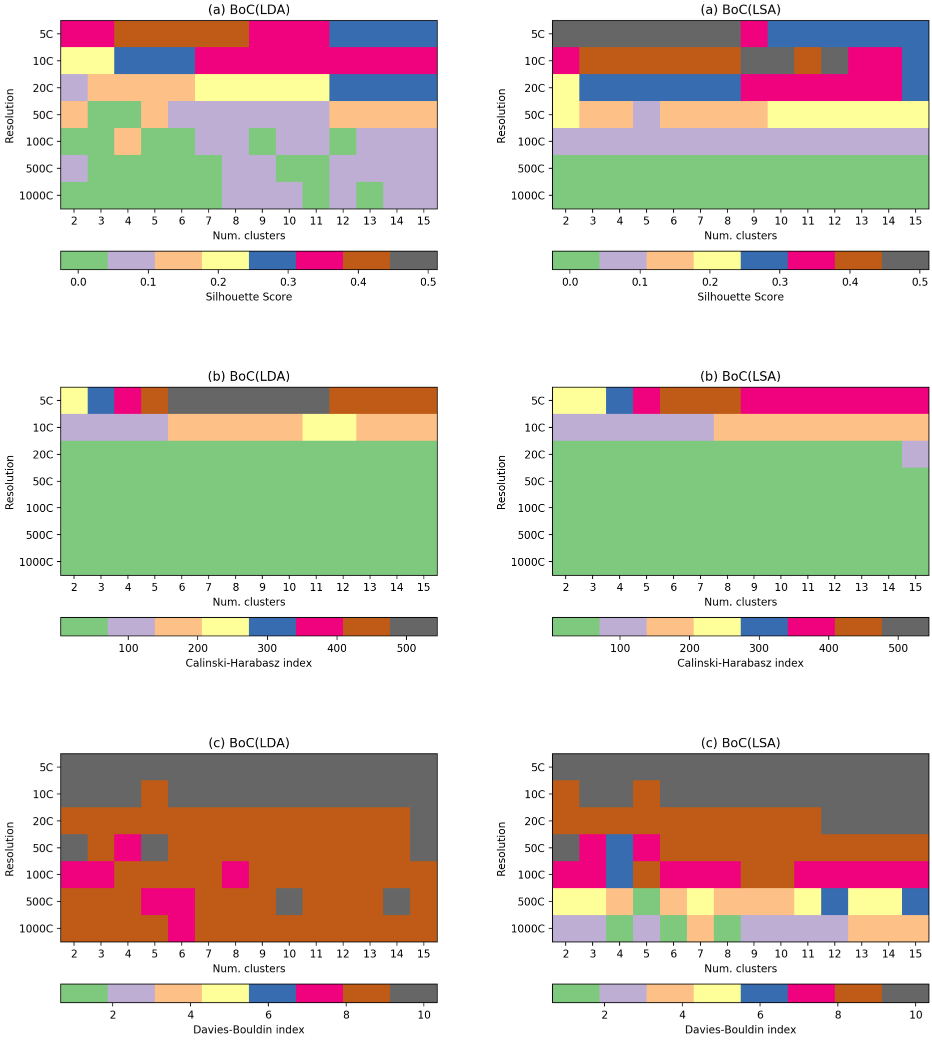

Figure 3: Heatmaps showing the impact of the resolution parameter (p) in the clustering task. First row depicts results in terms of the s score, second row shows the CH index, and third row represents the DB index. Graphs in the same column were generated using the same approach for inferring word representations, specifically, here we are comparing LDA and LSA approaches.

across the three evaluation metrics, although is more clear for the s and CH indexes; (2) inferring word representations with LDA and LSA (Figure 3) allows us to obtain better performance across different values of k. In general, these experiments indicate that low-resolution values (5C, 10C) are preferable for obtaining the best clustering performance in the n-tv dataset.

Additionally, we evaluated our proposed method in three benchmark datasets, namely: Reuters 8 (Cardoso-Cachopo, 2007), AG's News (Zhang et al., 2015), and 10KGNAD (Schabus et al., 2017). Table 3 shows the obtained results in terms of the s score (SH), and clustering accuracy (ACA) values. It is important to mention that although these three datasets are labeled, we cannot compute the traditional Accuracy as in a supervised classification task because the k-means will assign an arbitrary label to every cluster it forms. However, what we can do is to compute the Average Clustering Accuracy (ACA) measure, which gives the accuracy of the clustering no matter what the actual labeling of any cluster is, as long as the members of one cluster are together. Traditionally, for obtaining the ACA value it is necessary to figure out what is the best setting that would yield me the maximum clustering accuracy. For our performed experiments, we used the sklearn linear_assignmen function, which uses the Hungarian algorithm to solve this problem.

As can be observed in Table 3, Boc(LDA) experiments were performed only for 5 and 10 concepts. We do not report results with a higher number of concepts because the LDA approach was not able to obtain more than 10 topics with high probability distributions, in other words, for greater values than 10 the employed LDA implementation generated empty topics for all the three datasets.

The first four rows represent the considered baselines. As can be noticed, the CNN approach performs well in the AGs News and 10KGNAD dataset, while for the R8 dataset, the traditional BoW obtains a competitive performance. In general, we can conclude that using the LDA approach for inferring the underlying semantics represents the best approach for inferring efficient highly-dense concepts. The BoC(LDA-5C) and BoC(LDA-5C) configurations obtain good results in terms of SH and ACA metrics in the R8 and AGs News datasets respectively.

6.2. Overall performance

From the previous analysis, we choose $p = 5$ as the best resolution value, since in two out of the three considered metrics, when the number of concepts is equal 5 we obtain better performances. Therefore, the next set of experiments was done using this as the number of concepts[9] and we compare our proposed approach against baselines described in section 5.2. Figure 4 shows the obtained results across the three considered evaluation metrics. Contrary to the previous section, here we kept the

[9] Represented as the '-5C' suffix in the experiments.

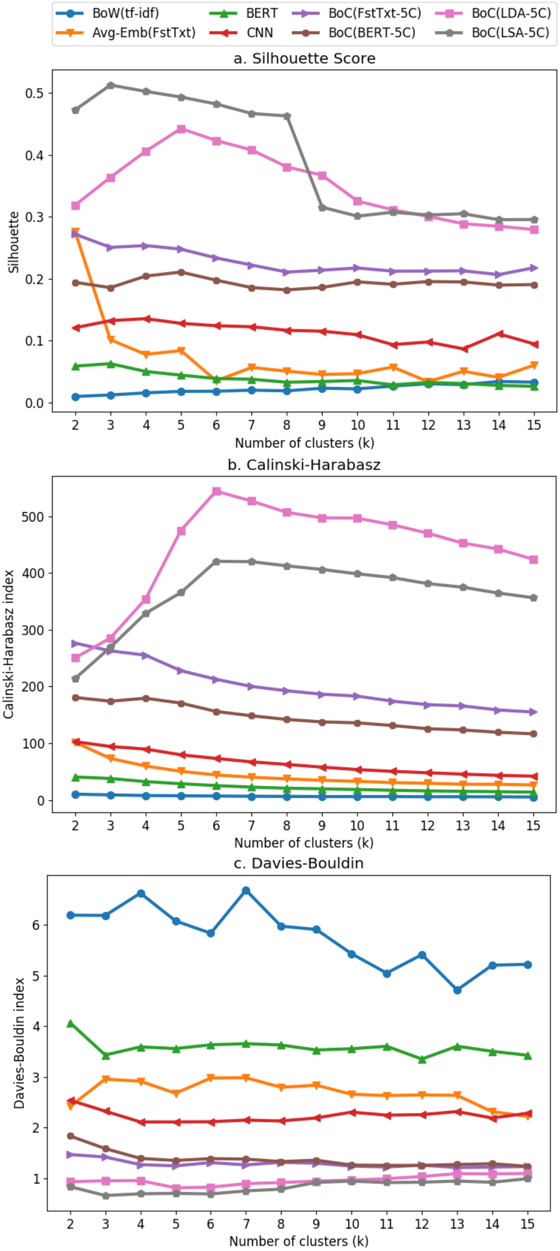

Figure 4: Clustering performance across several values of k: (a) s score, (b) CH index, and (c) DB index

Model	**R8**		**AGs News**		**10KGNAD**	
	SH	ACA	SH	ACA	SH	ACA
BOW	0.055	0.641	0.012	0.271	0.020	0.424
Avg-Emb(FstTxt)	0.054	0.474	0.042	0.409	0.223	0.225
BERT	0.077	0.378	0.041	0.599	0.039	0.368
CNN	0.079	0.407	0.057	0.623	0.158	**0.618**
BoC(FstTxt-5C)	0.279	0.312	0.300	0.361	0.221	0.327
BoC(FstTxt-10C)	0.199	0.325	0.203	0.344	0.231	0.513
BoC(FstTxt-20C)	0.131	0.322	0.158	0.403	0.185	0.503
BoC(FstTxt-50C)	0.098	0.319	0.122	0.579	0.116	0.495
BoC(FstTxt-100C)	0.088	0.364	0.086	0.539	0.073	0.485
BoC(FstTxt-500C)	0.057	0.392	0.043	0.610	0.034	0.527
BoC(FstTxt-1000C)	0.066	0.446	0.030	0.597	0.025	0.499
BoC(LSA-5C)	0.162	0.453	0.236	0.457	0.225	0.476
BoC(LSA-10C)	0.304	0.583	0.183	0.484	0.236	0.471
BoC(LSA-20C)	0.262	0.595	0.095	0.454	0.179	0.421
BoC(LSA-50C)	0.149	0.619	0.158	0.292	0.127	0.427
BoC(LSA-100C)	0.133	0.633	0.057	0.292	0.073	0.448
BoC(LSA-500C)	0.056	0.701	0.050	0.396	0.005	0.418
BoC(LSA-1000C)	0.085	0.592	-0.010	0.459	0.027	0.453
BoC(LDA-5C)	0.349	0.504	**0.424**	**0.793**	0.341	0.455
BoC(LDA-10C)	**0.388**	**0.721**	0.237	0.617	**0.384**	0.495

Table 3: Additional experiments on three benchmark datasets. Results are reported in terms of Silhouette score (SH), and average clustering accuracy (ACA).

original configuration of the DB index, i.e., the lower the obtained score, the better the performance of the clustering approach.

Notice that traditional BoW(tf-idf) and Avg-Emb(FstTxt) techniques obtain the worst performance overall. Similarly, the BERT approach, which represents each document using the produced encoded by the last hidden layer of the pre-trained model of BERT, obtains comparable results to those from the Avg-Emb(FstTxt) technique. Although the CNNs method (Xu et al., 2015) improves the performance of the three previous baselines, its obtained results are far from reaching those obtained with the different configurations of our proposed approach.

From these experiments, it becomes clearer that the proposed approach performs better when concepts are inferred using either LDA or LSA techniques. If we concentrate on the s score only, the best performance is obtained when using BoC(LSA-5C) at $k = 3$ ($s = 0.51$), which represents a relative improvement of 73% against the best baseline, i.e., the CNN approach. Similarly, if we observe the CH index, the best result is obtained with BoC(LDA-5C) at $k = 6$ ($CH = 544.19$), which represents a relative improvement of 81.1% against the best result of the CNN approach. And finally, in terms of the DB index, the best performance is obtained with BoC(LSA-5C) at $k = 3$ ($DB = 0.66$), which represents a relative improvement of 68% in comparison to the CNN approach. Hence, the main observations from this analysis are: (*1*) proposed approach consistently improves, across three different metrics, traditional clustering techniques as well as some more recent approaches based on deep NN; (2) LDA and LSA techniques allow inferring better word representations, improving clustering results in comparison to SOTA methods such as BERT encodings.

6.3. Manual evaluation

To judge the quality of the generated groups, we have taken a subset of 30 articles and performed a small manual annotation experiment using 6 human experts.

For this exercise, we randomly select 30 articles from the n-tv dataset. Every annotator was instructed to identify 5 different clusters, i.e., they had to organize the information into five semantically related groups. The only restriction given is that each group should have at least one document and the same document can not be assigned to more than one cluster. We choose 5 as the number of clusters to identify, as from the previous experiments (see Figure 4) we observed that with $k = 5$ as a middle point, it is possible to obtain good performance on all the considered metrics. We evaluated the annotator's agreement using the Kappa metric (Cohen, 1968). Resulting in a Kappa score of **0.49** which indicates a moderate agreement.

We performed a detailed analysis of the identified groups, and it was clear from the exercise that spotted topics were: ‘technology’, ‘economy’, ‘politics’, ‘car industry’, and ‘financial education’. We observed that annotators tend to disagree on the class of the document when the categories might be related to ‘economy’, ‘politics’, and ‘financial education’, similarly when a document might belong to ’technology’ and ’car industry’. However, using a majority vote scheme, we decided on the final class of each document, and we used these 30 documents as a test set. We evaluate our method using the BoC(LDA-5C) configuration, and we were able to obtain a **70%** accuracy in the classification process. In Figure 5 we show the clusters' visualization under this configuration.

7. Conclusions

In this paper, we proposed using highly dense representations, denominated low-resolution concepts, for clustering German broadcast media contents. The proposed approach infers the fundamental semantic elements contained in the input dataset, which are used for suggesting optimal clusters configuration. Performed experiments

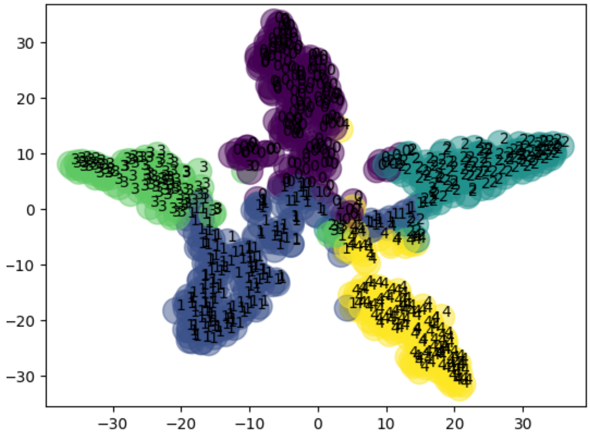

Figure 5: Formed clusters using the BoC(LDA-5C) configuration with $k = 5$. Found topics with the LDA approach are: *i*) chef (boss), autos (cars), deutschland (germany), zukunft (future), diesel (diesel); *ii*) euro (euro), prozent (percent), geld (money), experten (experts), deutschland (germany); *iii*) unternehmen (company), usa (USA), milliarden (billions), trump (Trump), eu (EU); *iv*) kunden (customers), google (Google), mitarbeiter (employees), online (online), facebook (facebook); and *v*) startup (sartup), deutschland (germany), daten (data), idee (idea), welt (world).

demonstrate that using small resolution values provides a better clustering performance, which is consistent across three different internal evaluation metrics, and in four different datasets. Particularly, the proposed framework is not dependent of any particular concise semantic analysis method for inferring concepts; however, when concepts are detected using the LDA and LSA approaches, the clustering performance tends to improve, obtaining relative improvements of 73%, 81%, and 68% under Silhouette, Calinski-Harabasz, and Davies-Bouldin indexes respectively. Finally, we would like to highlight one major advantage of our proposed approach, which is interpretability. As a result of the representation process, produced vectors are easy to interpret, facilitating end users understanding the found semantics and the decisions made by the system.

As future work, we plan to evaluate our proposed approach in similar datasets, i.e., very short texts, from a very narrow domain, and as the result of automatic transcription process from spontaneous speech. Is it possible to imagine, the latter represents a more challenging scenario since automatic transcription systems have many errors that might affect the performance of text-based methods.

Acknowledgments

We are very grateful to the annotators who help us manually validating the topics in the n-tv dataset: Alicia Illi, Noémi Stalder, Luca Schöb, Jasmin Staubli, Chaira Tremml, Angela Mürner. We also like to thank Daniele Zuccheri, and Jan Aeberli for providing access to the n-tv transcripts in the first place.

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

Bibliography

Adhikari, Ashutosh, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. *CoRR*, abs/1904.08398, 2019.

Blei, David M, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051.

Caliński, Tadeusz and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Cardoso-Cachopo, Ana. Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.

Cohen, Jacob. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. doi: 10.1037/h0026256.

Davies, David L and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

De Boom, Cedric, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, 2016. doi: 10.1016/j.patrec.2016.06.012.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Doulaty, M, O Saz, RWM Ng, and T Hain. Automatic Genre and Show Identification of Broadcast Media. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016. doi: 10.21437/Interspeech.2016-472.

Hinton, Geoffrey E and Russ R Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614, 2009.

Huang, Eric H, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proc. ACL*, pages 873–882, 2012.

Kim, Han Kyul, Hyunjoong Kim, and Sungzoon Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017. doi: 10.1016/j.neucom.2017.05.046.

Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Le, Quoc and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

Li, Chenliang, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174, 2016. doi: 10.1145/2911451.2911499.

Li, Zhixing, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441–448, 2011. doi: 10.1016/j.patrec.2010.11.001.

López-Monroy, Adrian Pastor, Fabio A González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225, 2018. doi: 10.18653/v1/N18-1110.

Miao, Yishu, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736, 2016.

Morchid, Mohamed and Georges Linarès. A LDA-based method for automatic tagging of Youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, 2013. doi: 10.1109/WIAMIS.2013.6616126.

Ostendorff, Malte, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. Enriching BERT with Knowledge Graph Embeddings for Document Classification, 2019.

Rendón, Eréndira, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5 (1):27–34, 2011.

Ribeiro-Neto, Berthier and Ricardo Baeza-Yates. Modern information retrieval. *Addison-Wesley*, 4:107–109, 1999.

Rousseeuw, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL `http://www.sciencedirect.com/science/article/pii/0377042787901257`.

Schabus, Dietmar, Marcin Skowron, and Martin Trapp. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan, August 2017. doi: 10.1145/3077136.3080711.

Sheri, Ahmad Muqeem, Muhammad Aasim Rafique, Malik Tahir Hassan, Khurum Nazir Junejo, and Moongu Jeon. Boosting Discrimination Information Based Document Clustering Using Consensus and Classification. *IEEE Access*, 7:78954–78962, 2019. doi: 10.1109/ACCESS.2019.2923462.

Silveira, Denys, Andr'e Carvalho, Marco Cristo, and Marie-Francine Moens. Topic modeling using variational auto-encoders with Gumbel-softmax and logistic-normal mixture distributions. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. doi: 10.1109/IJCNN.2018.8489778.

Staykovski, Todor, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. Dense vs. Sparse Representations for News Stream Clustering. In *Text2Story@ ECIR*, pages 47–52, 2019.

Teh, Yee W, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

Wang, Rui, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural Topic Modeling with Bidirectional Adversarial Training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 340–350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.32. URL `https://www.aclweb.org/anthology/2020.acl-main.32`.

Xu, Jiaming, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short Text Clustering via Convolutional Neural Networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, 2015. doi: 10.3115/v1/W15-1509.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

Address for correspondence:
Shantipriya Parida
`shantipriya.parida@idiap.ch`
Idiap Research Institute
Rue Marconi 19, 1920 Martigny
Switzerland.