IDIAP RESEARCH REPORT

CHALLENGES IN BROADCAST MEDIA CONTENT CATEGORIZATION

Shantipriya Parida[a] Esaú VILLATORO-TELLO[b]
Petr Motlicek

Idiap-RR-02-2021

APRIL 2021

[a]Idiap Research Institute
[b]Idiap

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

CHALLENGES IN BROADCAST MEDIA CONTENT CATEGORIZATION

Shantipriya Parida[⋆] *Esaú Villatoro-Tello*[⋆†] *Petr Motlicek*[⋆]

[⋆]Idiap Research Institute
Rue Marconi 19, 1920 Martigny, Switzerland
{firstname.lastname}@idiap.ch
[†]Language and Reasoning Research Group,
Universidad Autónoma Metropolitana, Cuajimalpa, Mexico.
evillatoro@correo.cua.uam.mx

ABSTRACT

With the sheer volume of content which the media companies need to create and distribute, the categorizing and classifying the content becomes a challenging task. The manual classification of media content is too time-consuming, inefficient, and expensive. There is an urgent need to automatically classify such data with high accuracy. In this paper, we propose an unsupervised approach to categorize the textual documents of broadcast media content. The proposed approach employs German language word embeddings in combination with *tf-idf* (term frequency-inverse document frequency) information for representing documents. The categories' names are derived from the found clusters based on their semantic relevance. The proposed approach is evaluated on a real-life dataset extracted from a German TV channel. Results indicate that using word embedding information is effective compared to the traditional *tf-idf* representation, obtaining a relative improvement of 17.61%. Additionally, we perform a qualitative analysis that addresses some of the challenges.

Index Terms— word embedding, spoken documents, document clustering, topic detection

1. INTRODUCTION

Video contents rapidly increase over the last few years, though it's difficult to exploit stored collections due to lack of structuring and reliable information associated with these contents [1, 2]. This represents one of the challenging scenarios for automatic categorization systems. Given that the manual classification of this type of data is expensive, there is an urgent need for automatic tools that can reliably classify these documents. Document categorization is one of the important tasks in the fields of machine learning, speech, and natural language processing [3]. Accordingly, the task of unsupervised document categorization is a fundamental research topic that aims at grouping similar objects into clusters based on their semantic similarities [4]. The similarity measure is an important factor in the clustering algorithm, and its efficiency depends on the feature representation space. The k-means clustering is one widely used partitioning algorithm, which employs the Euclidean distance in the feature space for dividing the objects into k clusters [5]. Distributed word representations or word-embeddings is a real-valued vector representation of words by embedding both semantic and syntactic meanings obtained from unlabeled large corpus [6] and have proven their efficiency in many NLP related tasks, such as sentiment analysis [7], question answering [8], author profiling [9], etc. Nevertheless, although word-embeddings provide a low-dimensionality and a non-sparse representation of documents, its pertinence, to the best of our knowledge, has not been evaluated on the analyzed task, i.e., categorizing short transcript texts from broadcast media content.

This paper proposes an unsupervised approach for the categorization and topic detection from broadcast media documents. The proposed approach combines the information extracted from German word embeddings with traditional *tf-idf* scores to effectively represent documents. Our main contribution is two-fold: (i) we propose a clustering approach based on multilingual word embeddings, and (ii) we evaluate our method on a real-life dataset of German spoken documents. We foresee this work will represent an important contribution to the development of novel methodologies in the field of unsupervised document clustering of broadcast media content.

The paper is organized as follows: Section 2 describes the proposed method used in this paper. Section 3 describes the dataset and its statistics. Section 4 presents the experimental setups used in our approach. Section 5 illustrates our results and our analysis followed with a conclusion in Section 6.

2. PROPOSED METHOD

The overall model of the proposed approach is shown in Figure 1, together with an overview provided by Algorithm 1.

The textual transcripts of the speech segments are the main input to our method. Once the transcripts ($\mathcal{D}$) of the speech data are obtained for each article, our method requires

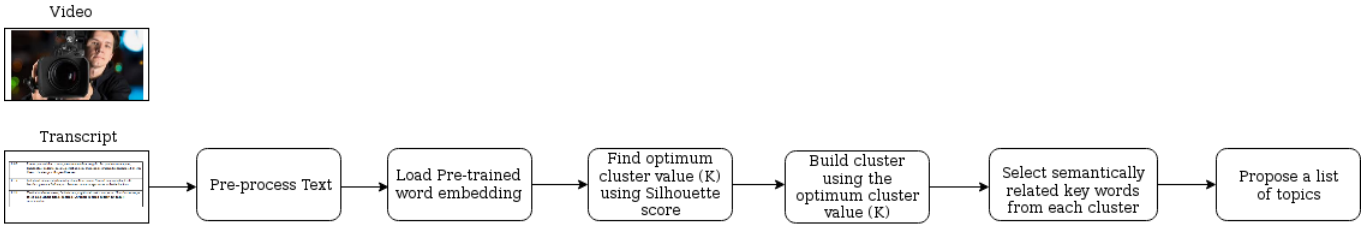

Fig. 1. Overall process of the proposed approach.

two additional parameters: (i) the number of required clusters k, and (ii) pre-trained word embeddings $\mathcal{G}$. As known, the goal of word embeddings is to capture semantic and syntactic regularities present in language from large unsupervised sets of documents. This type of approach takes as input a large corpus of texts and produces a vector space, typically of a few hundred dimensions, where each term in the corpus is assigned to a corresponding vector $\mathbf{w_i}$ in the space. Thus, once the word vectors are computed and positioned in the vectorial space, words that share common contexts in the corpus are located in close proximity to one another in the space [10]. In this paper, We used word embeddings trained with FastText[1] on 2 million German Wikipedia articles.

Overall, the proposed algorithm works as follows: The first step is to perform text normalization (in German) (i.e. removing stop-words, numbers, special symbols, and all the words are converted to lower-case). Then, for every word in the vocabulary, we compute its *tf-idf* score. The *tf-idf* weight is a score frequently used in information retrieval and text mining [11]. This weight is a statistical measure that indicates the importance of a word to the document in a corpus [12]. The document's representation ($\vec{\mathcal{D}}$) is built in the third step of the proposed algorithm. For this, we followed a popular strategy that consists of averaging term-vectors of words ($\mathbf{w_i}$) in each document [13]. However, unlike other methods, we consider the term's importance by multiplying the word vector by α_i, which represents the *tf-idf* score of the term i. After this step, we apply the k-means clustering algorithm, and retrieve the assigned labels $\mathcal{L}$ for each document, and the cluster's centroids μ. Finally, to obtain the topic name, we retrieve from the embeddings space $\mathcal{G}$ the closest words to every centroid μ_l. Obtained words represent the main topic for each of the k clusters.

Data: Document collection $\mathcal{D}$, number of required clusters k, German FastText embeddings $\mathcal{G}$
Result: Assigned labels $\mathcal{L}$ to each $d_j \in \mathcal{D}$, Set of proposed topics ($\mathcal{N}$) for clusters
1. $\mathcal{D}_p = pre\_processing(\mathcal{D})$;
2. $\alpha = Compute\_TF\_IDF(\mathcal{D}_p)$;
3. **for** $d_j \in \mathcal{D}_p$ **do**
   3.1 Build document ($\vec{\mathcal{D}}$) representation:
   $\vec{d_j} = \frac{\sum_{t_i \in d_j} \alpha_i \cdot \mathbf{w_i}}{|d_j|}$
end
4. Obtain assigned labels and cluster's centroids $(\mathcal{L}, \mu)$ from applying the k-means algorithm to $\vec{\mathcal{D}}$;
5. **for** $\mu_l \in \mu$ **do**
   5.1 $\mathcal{N}_l = closest\_words(\mu_l, \mathcal{G})$;
end

Algorithm 1: Pseudo-code of the proposed method for categorizing broadcast media documents, and suggesting topics names to identified clusters.

3. DATASET DESCRIPTION

The dataset used in our paper is from n-tv[2], 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 are the speech segments (audio data) with an average duration of 1.5 minutes where each recording has multiple speakers in a noisy environment. The other part of the dataset are the textual transcripts (German) associated with the speech segments. The dataset contains both labeled (topic) and unlabelled data. Each of the transcript files contains articles (text documents) spread across different topics. For performing our experiments, we have used the unlabeled set of articles, i.e., a total of 697 articles.

Table 1 shows some statistics from the employed dataset; before applying any pre-processing operation and after pre-processing. We compute the average number of tokens, vocabulary, and lexical richness in the dataset. Lexical richness (LR) is a value that indicates how the terms from the vocabulary are used within a text. LR is defined as the ratio between the vocabulary size and the number of tokens from a text ($LR = |V|/|T|$). Thus, a value close to 1 indicates a higher LR, which means vocabulary terms are used only once, while values near to 0 represent a higher number of tokens used more frequently (i.e., more repetitive). A couple of main observations can be done at this point. On the one hand, notice that individual articles are very short, on average 104.13 tokens with an average vocabulary of 85.18 words, resulting in a very high LR. This means that very few words are

[1]https://www.spinningbytes.com/resources/wordembeddings/
[2]https://www.n-tv.de/

	W/O Pre-processing	
	Average (σ)	Total
Tokens	234.68 ($\pm$ 124.45)	163,572
Vocabulary	161.79 ($\pm$ 51.92)	22078
LR	0.717 ($\pm$ 0.073)	0.134
	W/ Pre-processing	
	Average (σ)	Total
Tokens	104.13 ($\pm$ 52.84)	72,580
Vocabulary	85.18 ($\pm$ 31.02)	18,927
LR	0.840 ($\pm$ 0.077)	0.260

Table 1. Statistics of the German News Channel text data in terms of number of tokens, vocabulary and lexical richness.

repeated within one article, very few redundancies, making the categorization task more challenging. On the other hand, globally speaking, the complete dataset has a low LR (0.26), which indicates, to some extent, that information across articles is very repetitive, even though articles might belong to different domains (topics). Finally, we measure the coverage ratio of the German word embeddings into our dataset, resulting in a 91.45% of coverage. This means that from the 18927 terms contained in the vocabulary, only 1617 terms are not represented in the embeddings space.

To judge the quality of the data, we have taken a few labeled data and performed a small exercise (manual evaluation) using 3 human annotators. We have taken 23 articles and 3 human annotators and the actual labels (given by an expert) as another annotator with 9 pre-defined categories including "other" as a category in case annotators feel, the article in a revision does not fall in any of the given categories. We have evaluated the annotator's agreement using the Kappa metric [14]. We have obtained a Kappa score of **0.22** which interprets as fair agreement. We observed though, for a few of the articles, the human annotators are agreed upon the categories, however, the agreement differs with the expert annotator (actual labels). The Kappa score indicates the complexity of the analyzed task in this particular type of article.

4. EXPERIMENTAL SETUP

As it is known, the k-means clustering algorithm has an important parameter, the number of required clusters k. To know the optimal number of clusters, we ran the k-means algorithm on all the articles for a range of k values. Given that the number of existing topics within the data is unknown, supervised evaluation metrics such as *accuracy, F-score, precision* or *recall* can not be applied. Nevertheless, it is necessary to measure the quality of formed groups, aiming at determining the clustering tendencies of the data, i.e., distinguishing whether (or not) a random structure exists in the data [15].

Accordingly, as the evaluation measure, we employ the Silhouette coefficient [16], which combines the concepts of cohesion and separation for each point of clusters. On the one hand, the cluster's cohesion measures how closely the objects in a cluster are related among them. On the other hand, the cluster's separation measures how well a cluster is distinguished from other clusters. The silhouette coefficient value varies from -1 to 1. A negative silhouette value depicts an undesirable result, meaning a bad clustering; on the contrary, positive values define a better quality in the clustering result.

As a baseline we employed a traditional Bag-of-Words (BoW) representation using a *tf-idf* weighting scheme. BoW representation is a traditional baseline used in many text categorization tasks.

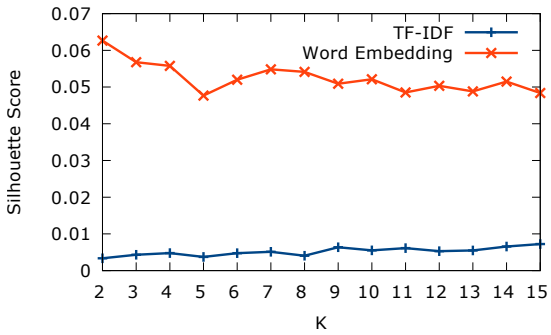

Fig. 2. Silhouette score analysis plot. The X-axis shows the value of k and Y-axis shows the Silhouette score. The Silhouette score based on *tf-idf* and word embedding are shown in different color labels.

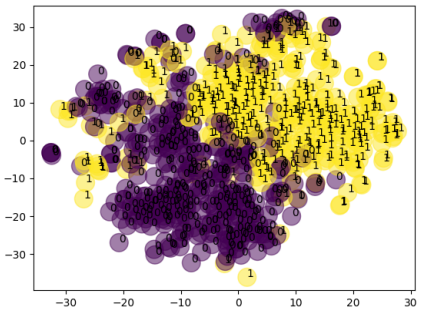

Fig. 3. Clustering view of total number of articles for two categories.

5. RESULTS AND DISCUSSION

Figure 2 shows the overall performance of our proposed method. As can be observed, our proposed approach consistently outperforms traditional *tf-idf* representation. Thus, for the following analysis, we have selected the value of $k = 2$ for which the Silhouette score is maximum [17], obtaining

a relative improvement of 17.61% compared to the baseline. Accordingly, the clusters' visualization is shown in Figure 3. From here, we took on the task of analyzing the formed clusters to determine their quality. For this, we compute the similarity matrix of the entire dataset preserving the order of the articles according to its cluster assignment; 354 articles are assigned to cluster 0, and 343 to cluster 1.

Therefore, in Figure 4, we show the results of the articles' similarity using the cosine distance metric. Each value in the X and Y axes represents an article, the first 343 belong to cluster 1, and the next 354 to cluster 0. As it is possible to observe, using a traditional BoW representation for comparing documents' similarity does not provide useful information. Under the BoW representation, very few documents from the same cluster have a clear similarity, i.e, very few dark dots. On the contrary, in Figure 5 we can see more clearly the formed clusters (darker squares on the upper-left and lower-right corners of the graph). Similarly to the analysis performed in Figure 4, we measure document similarity, but using the averaged word embeddings for representing each document. Finally, the inferred topics from the clusters are shown in the Table 2, which are visualized in the word embedding space in Figure 6. From these, we can observe that topic 0 (green dots) refers mainly to economic aspects such as investing, while for topic 1 (blue dots), it refers to consumers' concerns.

Cluster No.	Topic	Gloss
0	marketinganstrengungen	marketing efforts
0	informationsverpflichtungen	information obligations
0	wirtschaftspositionen	economic positions
0	verkaufsanstrengungen	selling efforts
0	investierenden	investing
1	verbraucherausstellungen	consumer exhibitions
1	realistischerweise	realistically
1	verbraucherausstellung	consumer exhibition
1	selbstverständilich	of course
1	ehrlicherweise	honestly

Table 2. Sample topics from the clusters. The two clusters are represented as "0" and "1" along with the topics and their gloss (translated for understanding).

6. CONCLUSIONS

This paper discusses the challenges that exist in categorizing contents extracted from the broadcast media. We proposed an unsupervised approach using German word embedding information, and showed the effectiveness of our method as compared to traditional *tf-idf* based approach while applying on the real media data. Our proposed approach can obtain a Silhouette score relative improvement of 17.61% in comparison to the baseline. Additionally, our proposed approach automatically infers the main topics from the clusters, thus providing useful information regarding the main content. Going forward, we plan to apply the technique on other language data to explore the generality of the proposed method.

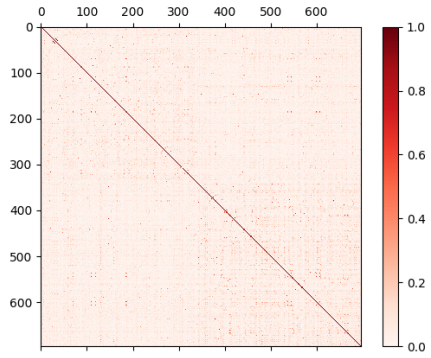

Fig. 4. Similarity matrix from obtained clusters using a traditional BoW with *tf-idf* weighting.

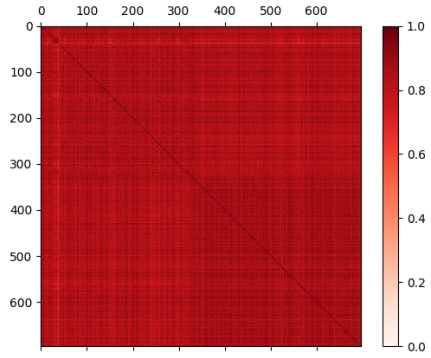

Fig. 5. Similarity matrix from obtained clusters using the FastText word embedding representation.

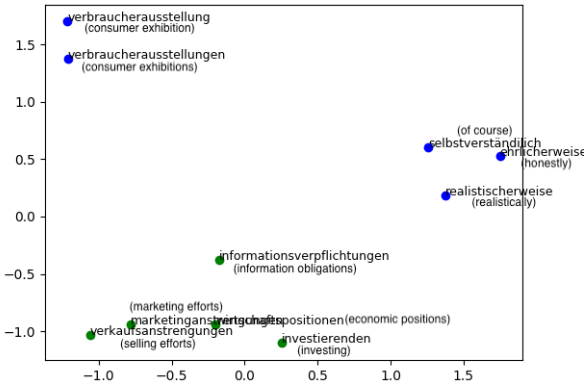

Fig. 6. Output word embedding. Topics from the two clusters are shown in different colors (blue and green).

7. REFERENCES

[1] Mohamed Morchid 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)*. IEEE, 2013, pp. 1–4.

[2] M Doulaty, 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.

[3] Naser SA Abu Sulaiman, "Automatic topic classification system of spoken arabic news," *Automatic Topic Classification System of Spoken Arabic News*, 2017.

[4] Christian Wartena and Rogier Brussee, "Topic detection by clustering keywords," in *2008 19th International Workshop on Database and Expert Systems Applications*. IEEE, 2008, pp. 54–58.

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[6] Bin Wang, Angela Wang, Fenxiao Chen, Yunchen Wang, and C-C Jay Kuo, "Evaluating word embedding models: Methods and experimental results," *arXiv preprint arXiv:1901.09785*, 2019.

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[9] Miguel A Álvarez-Carmona, Esaú Villatoro-Tello, Luis Villaseñor-Pineda, et al., "A comparative analysis of distributional term representations for author profiling in social media," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4857–4868, 2019.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al., *Modern information retrieval*, vol. 463, ACM press New York, 1999.

[12] Juan Ramos et al., "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*. Piscataway, NJ, 2003, vol. 242, pp. 133–142.

[13] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016.

[14] Jeroen Geertzen, "Inter-rater agreement with multiple raters and variables," *https://mlnl. net/jg/software/ira/, Retrieved May*, vol. 8, pp. 2014, 2012.

[15] Alba Núñez-Reyes, Esaú Villatoro-Tello, Gabriela Ramírez-de-la Rosa, and Christian Sánchez-Sánchez, "A compact representation for cross-domain short text clustering," in *Advances in Computational Intelligence*, Grigori Sidorov and Oscar Herrera-Alcántara, Eds., Cham, 2017, pp. 16–26, Springer International Publishing.

[16] Peter J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[17] Chunhui Yuan and Haitao Yang, "Research on k-value selection method of k-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.