

GraphTMT: Unsupervised Graph-based Topic Modeling from Video Transcripts

1st Jason Thies
Social Computing
Technical University of Munich
Munich, Germany
jason.thies@tum.de

1st Lukas Stappen
EIHW
University of Augsburg
Augsburg, Germany
stappen@ieee.org

2nd Gerhard Hagerer
Social Computing
Technical University of Munich
Munich, Germany
ghagerer@mytum.de

3rd Björn W. Schuller
GLAM
Imperial College London
London, United Kingdom
bjoern.schuller@imperial.ac.uk

4th Georg Groh
Social Computing
Technical University of Munich
Munich, Germany
grohg@mytum.de

Abstract—To unfold the tremendous amount of multimedia data uploaded daily to social media platforms, effective topic modeling techniques are needed. Existing work tends to apply topic models on written text datasets. In this paper, we propose a topic extractor on video transcripts. Exploiting neural word embeddings through graph-based clustering, we aim to improve usability and semantic coherence. Unlike most topic models, this approach works without knowing the true number of topics, which is important when no such assumption can or should be made. Experimental results on the real-life multimodal dataset MuSe-CaR demonstrates that our approach GraphTMT extracts coherent and meaningful topics and outperforms baseline methods. Furthermore, we successfully demonstrate the applicability of our approach on the popular Citysearch corpus.

Keywords—topic modeling, graph connectivity, transcripts, k-components, clustering

I. INTRODUCTION

Hundreds of hours of videos are uploaded to YouTube every minute, enabling studies in various fields of research. For example, educational information on cancer treatment [1] and hearing aids [2] are studied in health-care, the influence on election campaigns in social sciences [3], and large-scale multimodal sentiment in multimodal machine learning [4], [5], [6], [7], [8]. For these approaches, researchers closely examine the videos for collection, labelling, and analysis, whereby visual patterns and metadata, e. g. , authorship, can be exploited. Nowadays, also transcripts – automatically created by YouTube – are available [9]. Since text is the most meaningful modality to understand contextual information, effective computer-assisted text analysis methods are needed.

Topic models that structure information into theme distributions have existed for many years. It has been performed on a range of different texts, including online social network data [10], [11], [12], journals [13], and transcripts [14], [15].

Given a transcript snippet: “*It comes with four turbochargers on [and] has an aught [⇒ naught] to 62 [⇒ 60] time of just 5.2 seconds and [...]*”, a typical two-way topic modeling procedure *first*, extracts the aspect terms e. g. , “turbochargers”, *second*, clusters the aspects into coherent topic clusters e. g. , “motorisation” = {“turbochargers”, “engine”, ...}. Automatic transcripts, however, bring unique challenges. Transcripts often have errors like missing words (“and”), incorrect (“62” ⇒ “60”), and similar sounding words (“aught” ⇒ “naught”) due to erroneous speech-2-text processing.

Video transcripts are an emerging data domain, however, the explicit use for topic modeling is understudied [15], [14], [16]. To broaden the perspective on this medium more evaluation and new approaches are needed. Recently, graph connectivity showed promising results on extracting topic from news articles [17]. Compared to other methods [18], [10], [19], the number of expected topics does not have to be explicitly determined a priori. In addition, graph modelling research has gained momentum in several areas, such as text classification [20] and video retrieval [21].

In this work, we propose a *Graph-based Topic Modeling approach for Transcripts* (GraphTMT). For benchmarking, we base our evaluation on *a*) a problem-specific multimedia dataset of car reviews, MuSe-CaR [6], and *b*) the popular written-text dataset Citysearch [22]. MuSe-CaR is one of the largest state-of-the-art video datasets for multimodal sentiment analysis research, containing almost 40 hours of video footage and transcripts of car reviews. The reported word error rate of the automatic transcript is estimated around 28 % [6]. To the best of our knowledge, studies on topic extraction have only been conducted in a supervised fashion [23], [24], [25] on this corpus. Furthermore, Citysearch is utilised to evaluate the applicability of our approach to other datasets. It covers written reviews from restaurant visits and is often featured for the task of aspect and topic modeling in previous works [22], [26], [27].

¹JT and LS contributed equally to this work.

Our contributions are as follows: We propose a novel graph-based approach for topic modeling for the emerging use case of video transcripts. It is the first time, an unsupervised extraction model is applied to a large-scale, noisy MuSe-CaR dataset packed with typical mistakes of automatic speech-to-text. The performance is extensively benchmarked on this dataset against conventional methods. Here, the semantic consistency of the topics is evaluated by assessing a common coherence measure. Furthermore, for a more human-centred evaluation approach of the results and to determine the semantic validity, we conduct a structured word intrusion user study with 31 subjects. Finally, we evaluate the coherence of our approach on a standard topic modeling dataset of product reviews to assess the potential for other use cases. Our results show that GraphTMT outperforms conventional methods on the MuSe-CaR datasets. For reproducibility, this paper is adjoined with a public Git repository¹.

II. RELATED WORK

A. Word Vector Based Topic Models

Topic modeling is often performed by clustering natural language embeddings, grouping semantically similar words together to discover the semantic structure of the underlying corpus [28], [29], [30].

Curiskis [28] compared a traditional topic modeling based on Latent Dirichlet Allocation (LDA) with clustering embedding approaches. All models were applied to Twitter and Reddit textual data. His study indicated that weighted and unweighted embedding clustering has the potential to outperform traditional approaches when using word2vec.

Recently, Sahlgren [29] compared document-based topic modeling to word-based topic modeling. The word-based topic models used utilized embeddings for each prominent word, and the document-based model used document embeddings. The study showed that word-based topic modeling resulted in less or no overlap, more unique topics, and higher average topic coherence. Furthermore, Wang et al. [30] recently evaluated the performance of different topic modeling approaches on Twitter data, applying embedding clustering. The study indicates that more advanced models, such as BERT, do not necessarily outperform approaches on distributed embeddings.

B. Graph-based Topic Models

While these studies used clustering methods to create semantically related word groups, comparatively few have worked with graphs for topic extraction. This paper aims to motivate research in using graph connectivity for topic modeling. While common clustering techniques require strict hyperparameters, e.g., K-Means requires the true number of topics, K-Components [31] does not. Altuncu et al. [17] used graph connectivity and document embeddings to extract topics. The graph nodes represent documents, and the edges are weighted by the cosine similarity of the respective document

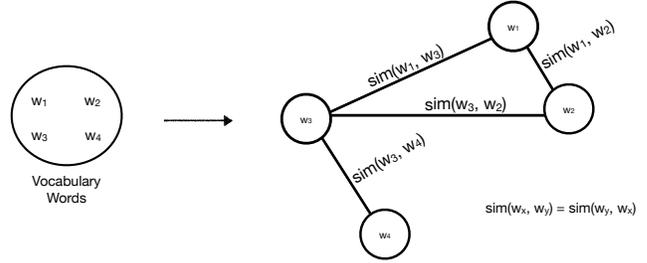


Fig. 1: Illustration of a word embedding graph. Each node represents a word from the vocabulary and each edge is weighted by the similarity between the adjacent nodes. The edges are undirected so that $\text{sim}(w_i, w_j) = \text{sim}(w_j, w_i)$.

pair. The study applied minimum spanning tree and community detection to extract document groups, representing the topics of the corpus. The study concluded that graph connectivity outperforms standard clustering techniques (e.g., K-Means). Graph-based clustering approaches have been successfully utilized in various applications, e.g. in crime pattern analysis [16] and cohesive subgraphs' discovery for social networks [32].

C. Topic Modeling on Video Transcripts

There are promising applications and use cases of topic modeling related approaches on YouTube video transcripts. Morchid and Linarès [15] used LDA-based topic modeling on self-generated YouTube video transcripts to improve automatic tagging of the uploaded videos. While the overall tagging robustness improved compared to conventional approaches, absolute performance in predicting user-provided tags remained low. The authors argued that this is due to subjectivity and high word error rate of their custom speech recognition system. More recent works are based on the video transcripts provided by YouTube itself. Basu et al. [14] apply preprocessing using automatic spell checking and irrelevant word removal. They utilize LDA for soft assignment of topics to teaching videos and texts. Furthermore, latent semantic indexing, a technique related to topic modeling, has been leveraged for search indexing on YouTube transcripts [33]. Despite existing topic modeling applications, to the authors' best knowledge, there are no coherence evaluations of topic modeling technology on YouTube transcripts. Such tool would be helpful to extract opinion targets for opinion mining purposes on video product reviews in an unsupervised manner [27], widely established approach on text-based product reviews. Our goal is to foster this research on publicly available video transcripts for market research purposes.

III. APPROACH: GRAPHTMT

In this section, we describe our proposed graph-based topic modeling approach. The ultimate goal of GraphTMT is to create and split a word embedding graph, into subgraphs based

¹Our code can be found at https://github.com/JaTrev/unsupervised_graph-based

on edge connectivity. The resulting subgraphs, similar to word embedding clusters, hold semantically related words and are considered the prominent topics of the corpus.

A. Word Embedding Graph

Given a set of vocabulary words W ($|W| = n$), a unique set of the most prominent corpus words, a word embedding graph $G = (N, E)$ is created consisting of $|N| \leq n$ nodes. Each node represents a vocabulary word and each undirected edge $e \in E$ is weighted by the cosine similarity score of the adjacent nodes (cf. Figure 1). Cosine similarity is used to represent the semantic similarity embodied within the trained embeddings [34]. A higher cosine score indicates higher semantic similarity, while an edge weighted with a low cosine score indicates that the adjacent words are not semantically related.

B. Edge Dropping

By weighting the edges, low-weighted edges can be removed from the graph without disconnecting subgraphs of high semantic similarity. To extract insightful topics from the graph, GraphTMT uses a percentile threshold p_t to remove low-weighted edges in E .

C. Graph-based Topic Modeling

Using the resulting (incomplete) graph, the k -component subgraphs [31], [35], [36] are calculated. A k -component is a maximal subgraph of the original graph having (at least) edge connectivity k , a minimum of k edges must be removed from such a k -component subgraph to split it into further subgraphs. These subgraphs are inherently hierarchical; a 1-connected graph can contain several 2-component subgraphs, each of which can contain multiple 3-component subgraphs. In Figure 1, $G_{sub} = (N_{sub}, E_{sub})$, with $N_{sub} = \{w_1, w_2, w_3\}$ and $E_{sub} = \{\text{sim}(w_1, w_3), \text{sim}(w_3, w_2), \text{sim}(w_1, w_2)\}$ is a 2-component subgraph of the given graph. Each k -component subgraph represents a topic discussed in the corpus. The top N representatives of each topic are selected based on node degree and node weights.

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluate our method on two real-world datasets. We focus on MuSe-CaR, applying different topic modeling approaches to the unique dataset but include the Citysearch corpus to demonstrate the applicability of GraphTMT outside of video transcripts.

a) MuSe-CaR: The MuSe-CaR [6] is a multimodal dataset gathered in-the-wild from English YouTube videos centred around car reviews. It was created with different computational tasks in mind, allowing researchers to improve the machine’s understanding of how sentiment and topics are connected. The in-the-wild aspect of MuSe-CaR refers to the natural conditions a video is captured in. It varies in recording equipment, recording setting, and soundscapes. The audio captures ambient noises (e.g., car noises), while the

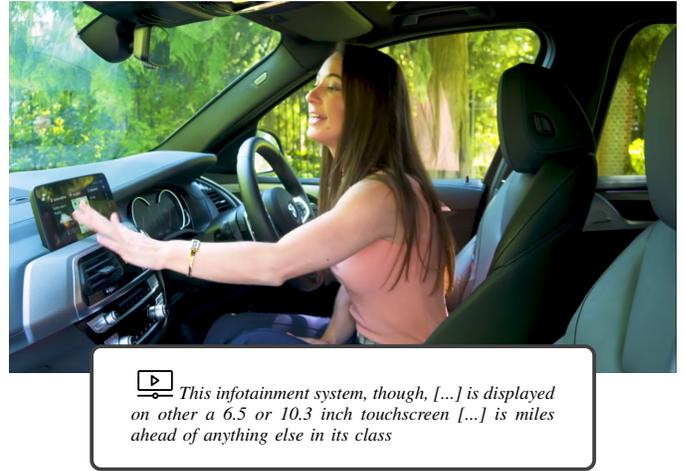


Fig. 2: Frame from MuSe-CaR (video id 2, 4:06) showing a User Experience segment and corresponding transcripts.

non-acted speech includes colloquialisms and domain-specific terms.

For our experiments, we use a preprocessed subset of the data featuring labelled topic segments², consisting of a total of 35h 39min of YouTube car review videos of approx. 90 speakers [23]. Consisting of real-life opinions about different aspects of modern vehicles, the dataset allows one to apply models to a large volume of user-generated data. The corpus includes 5 467 segments, each consisting of multiple sentences (total: > 20k sentences) with an average of 54 words. Long, encapsulated utterances are typical for transcripts. Video segments are assigned to one of ten topics: *Comfort, Costs, Exterior Features, General Information, Handling, Interior Features, Performance, Safety, Quality & Aesthetic, and User Experience*. The transcripts are generated by the authors using automatic Amazon Transcribe speech-to-text pipelines. Due to the in-the-wild factors, the error rate of the automatic transcripts is estimated to be relatively high and specified at around 28% with outliers of up to 39% on a subset of 10 hand-transcribed videos [6].

b) Citysearch corpus: Restaurant reviews from Citysearch³ have been widely used in previous works [22], [26], [27]. Citysearch was created in 2006. The project aims to provide a better understanding of patterns in user reviews and create tools to better analyse text reviews. The corpus contains over 50 000 restaurants reviews, written by over 30 000 distinct users. Ganu et al. [37] manually labelled a subset of 3 400 sentences using one of six topics: *Ambience, Anecdotes, Food, Miscellaneous, Price, and Staff*. The topic modeling approaches are evaluated based on this labeled subset.

B. Preprocessing

We begin by extracting the corpus vocabulary $W = \{w_1, w_2, \dots, w_n\}$ ($|W| = n$). The Natural Language Toolkit

²Download MuSe-Topic: <https://zenodo.org/record/4134733>

³Download Citysearch: <http://www.cs.cmu.edu/~mehrbod/RR/>, accessed on 29 April 2021

Fig. 3: Snippet from a review from the Citysearch corpus.

(NLTK) [38] part-of-speech (POS) tagger is used to collect POS tags for each word. Word tags have been successfully applied in previous studies [39], [16]. Stop word removal is applied to the Citysearch vocabulary, due to its larger size.

After extracting the corpus vocabulary W , we associate each word to a word embedding. The word2vec model [40] is used to learn these feature vectors, using the following parameters: window size = 15, epoch = 400, hierarchical softmax, and the skip-gram word2vec model [40]. For a fair comparison, this configuration is used in all settings.

Furthermore, we run preliminary experiments on MuSe-CaR and Citysearch utilising the POS tags (cf. Section III). The results indicated that using only nouns performs better on MuSe-CaR, regardless of the method, while the use of all parts-of-speech tags yields slightly better results on Citysearch (cf. Section VII) which we report in the following.

C. Baseline Approaches

Three baseline approaches are compared with GraphTMT: Latent Dirichlet allocation (LDA) [41], K-Means [42], and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[43].

LDA is a common topic modeling technique, using word co-occurrences to learn semantic clusters. It uses a Dirichlet prior on the topic distribution and the topic representatives distribution. LDA works with a bag-of-words (BOW) representation of the data. Each text is represented as a set of words and their cardinality, neglecting the sentence structure and context. Commonly, the BOW representation is translated into term frequency (TF) or TF-inverse document frequency (TF-IDF) matrix representation. K-Means is a common clustering technique used in topic modeling [18], [10], [19], [30], [17]. While LDA works on probability distributions of topics on the document, K-Means uses the distance between clusters. Similarly to LDA, K-Means commonly [10] uses the TF or TF-IDF matrix representation of the data. The algorithm simultaneously divides the dataset into a number of T_n clusters. The number of clusters is predefined, and the algorithm repeats two steps: an assignment and an update step. While in the assignment step, each data point is assigned to the cluster centroid based on the least squared Euclidean distance, the update step recalculates the centroids. HDBSCAN is a hierarchical and density-based clustering technique which creates a minimum spanning tree and condenses it into smaller trees to create clusters, stopping at C_{min} . Unlike K-Means, HDBSCAN allows for outliers.

Parameter	Values
Number of topics (T_n)	[4; 20]
Document-topic density (α)	[0.1, 0.4, 0.7, 1.0, $1/T_n$]
Word-topic density (β)	[0.1, 0.4, 0.7, 1.0]
Weighting strategy	[TF, TF-IDF]
Minimum cluster size (C_{min})	[5; 30]
Edge-connectivity (k)	[1, 2, 3]
Edge weight threshold (p_t)	[0.50, 0.60, 0.70, 0.80, 0.90, 0.95]

TABLE I: Parameter settings of the models

D. Measures

The different topic modeling approaches are measured by: (1) a coherence score, (2) intra-topic assessment, and (3) a user study.

a) (1) Coherence score: Topic coherence measures the degree of semantic similarity between topic representatives, the topic’s ten most eminent words. A model’s coherence score is the average of all topic scores. This study uses the c_v coherence score [44]. It is based on a sliding window approach that uses normalized pointwise mutual information (NPMI) and cosine similarity. Röder et al. [44] studied the correlation between numerous coherence scores and human judgement and found that c_v correlates best with human ratings.

b) (2) Intra-topic assessment: As coherence scores only capture the similarity between topic representatives, the intra-topic assessment compares the inferred topics with the dataset topic labels (the gold topics) [29]. It includes two measures:

- Topic coverage (T_c): how many gold topics are inferred? This is the proportion of gold topics that are included in the model’s topics. A larger number indicates better gold topic representation.
- Topic overlap (T_o): how much do the topics overlap? Each topic is given a label based on its representative, we compare these labels to find the proportion of duplicates. A small overlap indicates unique semantic structures.

c) (3) User study: Furthermore, a user study is conducted on MuSe-CaR models to measure the human interpretability of the inferred topics. Although topic coherence is measured, the interpretability of topics does not always align with coherence scores [45]. Our user study consists of the word intrusion task [46], [47], [45]. Each task is composed of six words, five representatives of a single topic, and a *not sure* option. The task is to find the word that represents a different topic, i. e., the intruder. Given the following intrusion task: {system, screen, diesel, menus, voice, entertainment, *not sure*}, all words besides “diesel” represent the same topic (infotainment). In this example, “diesel” is the intruder.

A models precision defines how well the intruder detected by the participants corresponds to the true intruder. We define the Word Intrusion Precision (WIP) by the fraction of subjects that find the correct intruders,

$$WIP_k^m = \sum_s \mathbb{1}(i_{k,s}^m = w_k^m) / S. \quad (1)$$

Let w_k^m be the intruder from the k^{th} topic inferred by model m and let $i_{k,s}^m$ be the intruder selected by participant s on the

Topic Models	T_n	c_v	T_C	T_O	WIP	NSF
LDA ($\alpha = 0.10, \beta = 0.70$)	8	.51	.60	.25	.43	.13
K-Means (TF-weighted)	8	.73	.60	.25	.61	.15
HDBSCAN ($C_{min}=6$)	11	.63	.60	.4	-	-
GraphTMT ($k = 1, p_t = 0.80$)	6	.76	.50	.17	.63	.08
GraphTMT ($k = 2, p_t = 0.80$)	5	.85	.40	.20	-	-
GraphTMT ($k = 3, p_t = 0.80$)	2	-	-	0	-	-

TABLE II: Results on MuSe-CaR for the different topic models and five different evaluation metrics: coherence score (c_v), topic coverage (T_C), topic overlap (T_O), WIP, and overall *not sure* fraction. Note, HDBSCAN was not included in the user study and only one GraphTMT model was assessed by the participants.

k^{th} topic. Let S denote the number of participants in the user study. Furthermore, the fraction of subjects that chose the *not sure* option (NSF) is captured.

To reduce study complexity, each model is assessed by half of its inferred topics (chosen at random) and each topic is assessed by a single word intrusion task. Overall the study includes 31 participants, each having an upper-intermediate English level (minimum of B2 in the Common European Framework of Language Reference).

V. MUSE-CAR EVALUATION

We first present the results on MuSe-CaR followed by the performance on Citysearch. All model parameters are optimized to maximize the topic model coherence. During the experimental process in this paper, adjustable parameters are set uniformly as shown in Table I. Any model inferring less than four topics and any topic with less than 5 representatives is not considered in our evaluation.

A. Coherence Score Comparison

In the first set of experiments, we compare our four models (LDA, K-Means, HDBSCAN, GraphTMT) on MuSe-CaR based on their coherence score. Table II shows the results of the best performing hyperparameters. Although the corpus has 10 gold topics, LDA and K-Means perform best with eight topics. The clustering-based model gets better scores using TF instead of TF-IDF. K-Means scores better than HDBSCAN but the hierarchical clustering techniques results in more topics. Our graph-based approach results in the highest coherence score ($c_v = .85$), achieving significant average topic coherence without specifying the number of topics (T_n) or the minimum size of a topic (C_{min}).

Furthermore, Table II shows the impact of k on GraphTMT. Increasing the edge connectivity parameter positively impacts the coherence score but at the expense of fewer topics. By increasing k , lower-weighted edges are removed from the graph, splitting or removing previously existing subgraphs. The new subgraphs only include the highest-weighted edges and most semantically related words. We note that GraphTMT ($k = 3$) results in only two topics, with ≥ 5 representatives, so it is not assessed in our experiments.

From these results, we can make the following observations: (1) the best performing approaches do not include 10 topics;

(2) baseline approaches can be used on MuSe-CaR to infer coherent topics; (3) clustering-based topic modeling achieves higher scores than probability-based LDA; (4) GraphTMT infers the most coherent topics without the need to specify the number of topics; and (5) by increasing k , the overall topic coherence of GraphTMT increases but T_n decreases.

B. Word Intrusion

As described in Section IV-D, the word intrusion task measures how well the inferred topics are interpretable by humans. Table II lists the precision results for the three best performing models (LDA, K-Means, GraphTMT) on MuSe-CaR. In our case, the c_v score aligns well with human judgement [44]. The best scoring topic model (GraphTMT) has the highest precision and the worst scoring model (LDA) has the lowest precision. Furthermore, GraphTMT has the lowest NSF score. These findings suggest that GraphTMT results in the most interpretable topics, underlining previous coherence results.

C. Intra-Topic Assessment

The previous two sections show K-Means and GraphTMT having the best topic coherence and WIP. This section looks at these two models' topic coverage and overlap (cf. Table II). K-Means has higher topic coverage than GraphTMT, but GraphTMT has a lower overlap between its topics. The overlap between topics reduces when we increase the edge connectivity constraint (k) but at the expense of topic coverage.

The eight topics inferred by K-Means (TF-weighted) are listed in Table III. Each topic is given a label, based on its topic representatives, and assigned to a gold topic. Overall, six unique gold topics can be matched ($T_c = 6/10$) but two topics are duplicates ($T_o = 2/8$).

Table III (middle) lists the six GraphTMT ($k=1$) topics. The topics include five gold topics ($T_c = 5/10$) and one overlap ($T_o = 1/6$). These topics can be compared to GraphTMT ($k=2$) in Table III. By increasing k , one of the two inferred *Infotainment* topics is removed from the graph, while *Performance* is split into two separate topics. Furthermore, the *Handling* topic was removed. As the coherence score increases with k , topics remaining in Table III (GraphTMT, $k = 2$) have a higher topic coherence score than the ones removed.

VI. CITYSEARCH EVALUATION

In the second part of our evaluation, we compare the performance of all four models on the Citysearch to show GraphTMT's applicability outside of YouTube transcripts. The models are compared on their coherence score, topic coverage, and topic overlap.

A. Coherence Score Comparison

Table IV lists the results of the best performing models based on their coherence scores. K-Means and GraphTMT ($k=3$) result in the highest coherence score, and LDA has the lowest. Similar to MuSe-CaR, K-Means gets better scores using TF instead of TF-IDF and increasing k has a positive effect

Inferred Topic	Topic Representatives	Gold Topic
K-Means		
<i>Handling</i>	suspension, handling, dampers, corners, chassis	Handling
<i>Infotainment</i>	menus, satnav, swivel, commands, entertainment	User Experience
<i>Interior Features</i>	dash, design, events, wood, plastic	Interior Features
<i>Performance</i>	engine, turbo, litre, cylinder, engines	Performance
<i>Safety</i>	detection, assist, safety, collision, airbags	Safety
<i>Storage</i>	storage, items, space, boot, hooks	General Information
<i>YouTube</i>	please, enjoy, click, share, wow	General Information
<i>Miscellaneous</i>	cars, guys, opportunity, brand, tomorrow	General Information
GraphTMT (K= 1)		
<i>Infotainment</i>	navigation, controls, touch, apple, buttons	User Experience
<i>Infotainment</i>	hand, pop, screen, entertainment, information	User Experience
<i>Passenger Space</i>	area, head, roof, room, headroom	Interior Features
<i>Handling</i>	suspension, corners, steering, gear, response	Handling
<i>Performance</i>	seconds, turbo, twin, acceleration, cylinder	Performance
<i>YouTube</i>	channel, dot, please, thanks, share	General Information
GraphTMT (k = 2)		
<i>Infotainment</i>	hand, pop, screen, entertainment, information	Infotainment
<i>Passenger Space</i>	seat, back, headroom, room, head	Handling
<i>Performance</i>	seconds petrol miles diesel economy gallon fuel	Performance
<i>Performance</i>	seconds, turbo, acceleration, twin, cylinder	Performance
<i>YouTube</i>	dot, channel, please, wow, share	General Information

TABLE III: List of topics extracted on MuSe-CaR where K-Means uses TF-weighted; GraphTMT uses $p_t = 0.8$.

Topic Models	T_n	c_v	T_c	T_o
LDA($\alpha = 1/T_n, \beta = 0.40$)	8	.48	.67	.50
K-Means(TF-weighted)	8	.64	.83	.38
HDBSCAN($C_{min}=5$)	3	.61	.33	.33
GraphTMT ($k = 1, p_t= 0.80$)	9	.40	.67	.56
GraphTMT ($k = 2, p_t= 0.80$)	6	.60	.67	.33
GraphTMT ($k = 3, p_t= 0.80$)	5	.64	.67	.20

TABLE IV: Results on the Citysearch for four different topic models (LDA, K-Means, HDBSCAN, GraphTMT) and three metrics: coherence score (c_v), topic coverage (T_c), and topic overlap (T_o).

on the coherence score of GraphTMT but reduces the number of topics. Citysearch has six gold topics, but K-Means infers eight and GraphTMT ($k=3$) results in five topics. At $k = 1$ our approach infers nine topics but has a lower score than LDA. HDBSCAN performed similar to K-Means but infers only three topics.

These scores show that our approach is applicable outside of YouTube transcripts, achieving the highest c_v score. Furthermore, they confirm a previous finding, increasing k results in a better score but fewer topics.

B. Intra-Topic Assessment

The previous scores show that K-Means and GraphTMT ($k=3$) have the best overall topic coherence. In the following, we look at their topic coverage and overlap (cf. Table II). Table V lists all K-Means topics, their inferred labels, and the model’s gold topic coverage. The table shows that K-Means covers five of the six gold topics ($T_c = .83$): *Ambience*, *Anecdotes*, *Food*, *Miscellaneous*, *Price*, but *Anecdotes*, and *Food* are captured twice ($T_o = .375$).

All GraphTMT models cover four of the six gold topics but as k increases, the topic overlap decreases. Table V lists the nine GraphTMT ($k=1$) topics, their inferred labels, and the topic coverage. Comparing these topics with the topics at $k=3$ shows the effect of k on GraphTMT. Increasing the edge connectivity parameter lowers the number of topics but can also let new topics turn up (i. e., *Ambient* is in GraphTMT

Inferred Topic	Topic Representatives	Gold Topic
K-Means		
<i>Ambience</i>	comfy, spacious, calm, sleek, couch	Ambience
<i>Miscellaneous</i>	appear, control, clue, sight, fooled	Miscellaneous
<i>Anecdotes</i>	yesterday, today, tonight, march, celebrate	Anecdotes
<i>Anecdotes</i>	refused, proceeded, busboy, ignored, annoyed	Anecdotes
<i>Price</i>	normal, pay, normally, expensive, afford	Price
<i>Location</i>	south, astoria, williamsburg, ues, houston	Miscellaneous
<i>Food</i>	yogurt, pear, pate, walnut, cinnamon	Food
<i>Food</i>	sliced, char, pate, prawn, chorizo	Food
GraphTMT (k = 1)		
<i>Food</i>	pickled, seed, puree, fennel, curried	Food
<i>Food</i>	poivre, hanger, hangar, flank, frites	Food
<i>Service (neg.)</i>	unhelpful, unattentive, unapologetic, arrogant, unfriendly	Staff
<i>Service (pos.)</i>	responsive, cordial, polite, gracious, professional	Staff
<i>Location</i>	washington, seaport, murray, madison, greene	Miscellaneous
<i>Location</i>	chelsea, downtown, soho, meatpacking, tribeca	Miscellaneous
<i>Location</i>	brand, england, yorker, orleans, yorkers	Miscellaneous
<i>Anecdotes</i>	incredible, outstanding, terrific, excellent, fantastic	Anecdote
<i>Time</i>	tuesday, wednesday, monday, friday, thursday	Anecdote
GraphTMT (k = 3)		
<i>Food</i>	pickled, seed, puree, fennel, curried	Food
<i>Service (neg.)</i>	unhelpful, unattentive, unapologetic, arrogant, unfriendly	Staff
<i>Service (pos.)</i>	responsive, engaging, sincere, caring, hospitable	Staff
<i>Anecdotes</i>	flavorless, tasteless, overcooked, undercooked, inedible	Anecdotes
<i>Ambient</i>	painted, tile, lantern, banquette, chandelier	Ambient

TABLE V: List of topics extracted on Citysearch where K-Means uses TF-weighted; GraphTMT uses $p_t = 0.8$.

($k=3$) but not in GraphTMT ($k=1$). This shows that topics can hold more semantics than indicated by their representatives, and increasing k can split an existing topic into semantically different topics, showing the hierarchical structure of our graph-based approach.

VII. DISCUSSION

We evaluated the competitiveness of our novel graph-based topic modeling approach to common alternatives (LDA, K-Means, HDBSCAN) on two different datasets (MuSe-CaR, Citysearch). Our experiments have shown that GraphTMT achieves the highest coherence scores on MuSe-CaR and Citysearch. Furthermore, the model’s edge-connectivity parameter (k) positively affects the coherence score but decreases the number of topics. These findings suggest that by varying k we can remove incoherent topics and words that do not semantically align with a topic. We should note that K-Means had the same coherence score on Citysearch but with more topics. All other models (LDA, HDBSCAN) scored less on both datasets. Although K-Means achieved a comparable score on Citysearch with more topics, the model requires one to predefine the number of topics. Since GraphTMT does not require a specification of the (true) number of topics, it is a good alternative if this information is not available, should not be predetermined, or a search for a suitable parameter k can not be performed. Moreover, the automatic retrieval of k by techniques such as the elbow method is controversial and rarely optimal [48].

In addition to comparing the semantic coherence of topics, we conducted a user study to assess the human interpretability of the MuSe-CaR topics. The study included the models with the highest coherence scores (LDA, K-Means, GraphTMT). As in previous studies, the resulting coherence scores align with the coherence scores [44], GraphTMT topics were more interpretable than topics from K-Means and LDA.

The intra-topic assessment allowed us to compare topics

from K-Means and GraphTMT, the two highest scoring models on both datasets. K-Means covered more gold topics, but GraphTMT resulted in topics with less overlap. Note that varying k revealed the hierarchical structure of GraphTMT, increasing the parameter can split a topic into two semantically different topics.

These findings suggest that GraphTMT provides a valid alternative to common topic model techniques as users can interpret the topics better, more unique topics are extracted, and the approach does not require the true number of topics. Overall, this study has shown the relevance of graph connectivity in topic modeling on two different datasets (YouTube transcripts and online restaurant reviews).

In our experiments, GraphTMT has proven to be very robust on a spoken dataset with a high word error rate. We want to validate these findings on other datasets in future work. Furthermore, we want to evaluate different preprocessing approaches for transcript. Another future aim is to compare different graph connectivity algorithms (e.g., clique percolation method) to find and develop even more effective approaches for topic extraction.

VIII. CONCLUSION

In this paper, we demonstrated the capability of graph-based topic modeling on real-world YouTube transcribed data (MuSe-CaR) and textual reviews (Citysearch). On the MuSe-CaR dataset, our proposed novel GraphTMT outperforms all three baseline models in terms of cluster coherence, uniqueness, and interpretability. An accompanying user study assessed the last one. On the Citysearch dataset, our method achieves competitive results to K-Means. However, the clusters produced by GraphTMT have less semantic overlap. We conclude that graph-based clustering is a valid alternative for topic modeling on transcripts and provides meaningful results on real-world text datasets. For the future, we will focus on an integrated approach of several modalities, such as, vision, audio and metadata as any attempt at drawing meaning from YouTube must consider all aspects.

REFERENCES

- [1] Corey Basch, Anthony Menafro, Jen Mongiovi, Grace Clarke Hillyer, and Charles Basch. A content analysis of youtube videos related to prostate cancer. *American Journal of Men's Health (AJMH)*, 2017.
- [2] Vinaya Manchaiah, Monica L Bellon-Harn, Marcella Michaels, and Eldré W Beukes. A content analysis of youtube videos related to hearing aids. *Journal of the American Academy of Audiology*, 2020.
- [3] Vassia Gueorguieva. Voters, myspace, and youtube: The impact of alternative communication channels on the 2006 election cycle and beyond. *Social Science Computer Review*, 2008.
- [4] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 2013.
- [5] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, 2011.
- [6] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, 2021.
- [7] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, page 5–14, New York, NY, USA, 2021. Association for Computing Machinery.
- [8] Lukas Stappen, Alice Baird, Michelle Lienhart, Annalena Bätz, and Björn Schuller. An estimation of online video user engagement from features of continuous emotions. *arXiv preprint arXiv:2105.01633*, 2021.
- [9] Ken Harrenstien. Automatic captions in youtube, Nov 2009. accessed on 29. April 2021.
- [10] Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034, 2020.
- [11] Paolo Missier, Alexander Romanovsky, Tudor Miu, Atinder Pal, Michael Daniilakis, Alessandro Garcia, Diego Cedrim, and Leonardo da Silva Sousa. Tracking dengue epidemics using twitter content classification and topic modelling. In *Current Trends in Web Engineering*. Springer International Publishing, 2016.
- [12] Dr. Rajesh Prabhakar Kaila and Dr. A. V. Krishna Prasad. Informational flow on twitter–corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 2020.
- [13] Carina Jacobi, Wouter Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 2015.
- [14] Subhasree Basu, Yi Yu, and Roger Zimmermann. Fuzzy clustering of lecture videos based on topic modeling. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2016.
- [15] 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.
- [16] Priyanka Das, Asit Kumar Das, Janmenjoy Nayak, Danilo Pelusi, and Weiping Ding. A graph based clustering approach for relation extraction from crime data. *IEEE Access*, 2019.
- [17] M. Tarik Altuncu, Sophia N. Yaliraki, and Mauricio Barahona. Graph-based topic extraction from vector embeddings of text documents: Application to a corpus of news articles. In *Complex Networks & Their Applications IX*. Springer International Publishing, 2021.
- [18] Suzanna Sia, Ayush Dalmia, and Sabrina Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2020.
- [19] Robert-George Radu, Iulia-Maria Rădulescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Mariana Mocanu. Clustering documents using the document to vector model for dimensionality reduction. In *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*. IEEE, 2020.
- [20] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.
- [21] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [22] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*. ACM, 2013.
- [23] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchun Du, Felix Hafner, Lea Schumann, Adria Mallo-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. ACM, 2020.
- [24] Lukas Stappen, Alice Baird, Erik Cambria, and Björn W Schuller. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*, 2021.
- [25] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W. Schuller. muse-toolbox: the multimodal sentiment analysis

- continuous annotation fusion and discrete class transformation toolbox. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, MuSe '21, page 75–82, New York, NY, USA, 2021. Association for Computing Machinery.
- [26] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 2010.
- [27] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. ACL, 2017.
- [28] Stephan Curiskis, Barry Drake, Thomas Osborn, and Paul Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 2019.
- [29] Magnus Sahlgren. Rethinking topic modelling: From document-space to term-space. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*. ACL, 2020.
- [30] Lili Wang, Chongyang Gao, Jason Wei, Weicheng Ma, Ruibo Liu, and Soroush Vosoughi. An empirical survey of unsupervised text representation methods on Twitter data. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. ACL, 2020.
- [31] David W. Matula. k-components, clusters, and slicings in graphs. *SIAM Journal on Applied Mathematics*, 1972.
- [32] Rong-Hua Li, Lu Qin, Jeffrey Xu Yu, and Rui Mao. Influential community search in large networks. *Proceedings of the VLDB Endowment (PVLDB)*, 2015.
- [33] Diana Iulia Bleoancă, Stella Heras, Javier Palanca, Vicente Julian, and Marian Cristian Mihăescu. Lsi based mechanism for educational videos retrieval by transcripts processing. In Cesar Analide, Paulo Novais, David Camacho, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2020*, pages 88–100, Cham, 2020. Springer International Publishing.
- [34] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics (TACL)*, 2015.
- [35] Jordi Torrents and Fabrizio Ferraro. Structural cohesion: Visualization and heuristics for fast computation. *Journal of Social Structure (JoSS)*, 2015.
- [36] Douglas R White and Mark Newman. Fast approximation algorithms for finding node-independent paths in networks. *SSRN Electronic Journal*, 2001.
- [37] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases*. ACM, 2009.
- [38] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.
- [39] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE International Conference on Data Mining (ICDM)*. IEEE, 2003.
- [40] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR)*. ACL, 2013.
- [41] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 2003.
- [42] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967.
- [43] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017.
- [44] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2015.
- [45] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 2009.
- [46] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, 2014.
- [47] Fiona Martin and Mark Johnson. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*. Australasian Language Technology Association (ATLA), 2015.
- [48] David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 1996.