Weakly Supervised Short Text Classification for Characterising Video Segments

Hao Zhang, Abrar Mohammed and Vania Dimitrova
School of Computing, University of Leeds, U.K.

Keywords: Video-Based Learning, Weakly Supervised Text Classification, Large Language Model.

Abstract: In this age of life-wide learning, video-based learning has increasingly become a crucial method of education. However, the challenge lies in watching numerous videos and connecting key points from these videos with relevant study domains. This requires video characterization. Existing research on video characterization focuses on manual or automatic methods. These methods either require substantial human resources (experts to identify domain related videos and domain related areas in the videos) or rely on learner input (by relating video parts to their learning), often overlooking the assessment of their effectiveness in aiding learning. Manual methods are subjective, prone to errors and time consuming. Automatic supervised methods require training data which in many cases is unavailable. In this paper we propose a weakly supervised method that utilizes concepts from an ontology to guide models in thematically classifying and characterising video segments. Our research is concentrated in the health domain, conducting experiments with several models, including the large language model GPT-4. The results indicate that CorEx significantly outperforms other models, while GLDA and Guided BERTopic show limitations in this task. Although GPT-4 demonstrates consistent performance, it still falls behind CorEx. This study offers an innovative perspective in video-based learning, especially in automating the detection of learning themes in video content.

1 INTRODUCTION

Globalization and rapid technological advancements have reshaped societal expectations, emphasizing the critical role of education. This requires a broad set of skills and learning contexts. The Council of Europe developed a Framework of Competences for Democratic Culture (Barrett, 2018) which categorizes the key competencies into into values, attitudes, skills, and knowledge and critical understanding—each vital for the cultivation of well-rounded individuals in a rapidly changing world. To meet the demands of democratic societies for future competences, learning throughout life and a shift to new learning paradigms are imperative. Crucially, life-wide learning has become the central paradigm for future education, as it accommodates learning across various stages and domains of life (Redeker et al., 2012).

Life-wide learning emphasizes the diversity of learning environments, suggesting that learning can unfold at any stage and across various life domains. This paradigm not only highlights the significance of formal, informal, and incidental learning experiences in diverse settings like home, leisure, community, and work, but also emphasizes their interconnectedness and complementarity (Commission of the European Communities, 2000). Yet, traditional educational systems, mainly set within fixed time frames and singular environments, are often ill-equipped to accommodate this holistic approach due to inherent limitations.

With the widespread accessibility of portable devices, the surge in internet users, and the exponential growth of freely available content, video technology's application in education has significantly expanded. This evolution allows learners to access numerous instructional videos at their convenience, fostering an environment conducive to life-wide learning (Sablić et al., 2020). However, learning from these videos demands not only a significant time investment for watching but also poses challenges for learners in identifying key points and linking them with relevant study domains (Bywater et al., 2021; Schlotterbeck et al., 2021). Addressing this need, effective characterization of videos becomes crucial, providing learners with a structured overview for efficient content navigation and targeted learning.

In this work, we address the unique challenges of characterizing video segments. The transcripts of video segments, often short and dense, present limitations to traditional supervised, unsupervised, and deep learning methods. To overcome these challenges and enhance the process of video segment classifica-

Zhang, H., Mohammed, A. and Dimitrova, V.
Weakly Supervised Short Text Classification for Characterising Video Segments.
DOI: 10.5220/0012618600003693

In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 2*, pages 197-204
ISBN: 978-989-758-697-2; ISSN: 2184-5026

tion for video segment characterisation, we propose a unique approach. It integrates the concepts from an ontology as weak supervision signals, capitalizing on the structured nature of an ontology to offer contextually relevant guidance. To assess the effectiveness of our approach, we conducted experiments involving several models, including the large language model (LLM) GPT-4. This paper presents a comprehensive analysis of the performance of different models in a health domain. It shows good performance of the CorEx model and highlights the limitations of GLDA, Guided BERTopic, and GPT-4 in this task. The approach can be used to automate the characterization of video content to support life-wide learning.

2 RELATED WORK

In this section, we briefly introduce related work from four different perspectives: video-based learning, short text classification, weakly supervised short text classification, and use of large language models.

2.1 Video-Based Learning

In the context of life-wide learning, video-based learning has emerged as a crucial educational modality. It provides learners flexibility, allowing them to access a wide range of learning resources anytime and anywhere, catering to their evolving democratic societies needs. However, video-based learning methods also bring new challenges to learners (see Section 1). This prompts us to consider how to characterise these videos effectively. Currently, methods for video characterisation primarily fall into two categories: manual characterisation and automatic characterisation. In the following, we will review these two approaches.

- **Manual Characterisation.** A common manual approach to video characterization involves note-taking, where essential information from the video is recorded and summarized (Dodson et al., 2019). However, this method heavily relies on human resources, making it resource-intensive and susceptible to errors. Another study proposes enhancing learner engagement and efficiency by having teachers emphasize crucial content using phrases, keywords, or questions (Tseng, 2021). Although aligned with the teacher's learning objectives, this approach is subjective and demands a significant amount of annotation work. Both methods face challenges in scaling with the abundance of online videos.
- **Automatic Characterisation.** Recent methods for automatic video characterization have primarily relied on two approaches: learner interaction and video content. In the learner interaction approach, one study propose using learner interactions and comments to identify key points that capture learners' attention in videos, aiming to characterize the video (Mitrovic et al., 2016). However, this approach depends on learners' responses and perspectives, and some may not effectively capture the video's key points. In the video content approach, another research utilizes slide content and teacher notes to characterize videos and segment them based on transitions between slides (Luca et al., 2019). However, this method primarily assesses the performance of its segmentation algorithm and overlooks whether its characterizations effectively aid learning.

These manual and automatic methods have their own limitations. For example, they often require significant human resources, and their effectiveness in assisting learning may not be comprehensive or well-evaluated. Video content classification is also one of the methods for video characterisation. Next, we will review various classification methods.

2.2 Short Text Classification

To achieve more granular video characterization, this research focuses on video segments. These segments have limited duration and the transcripts extracted from them are relatively short. These transcripts, categorized as 'Short Text', exhibit several distinct characteristics (Song et al., 2014; Li et al., 2017):

- **Sparsity.** The limited length of short texts leads to a significantly condensed feature space, impeding the extraction of effective linguistic attributes.
- **Ambiguity.** Short texts lack the comprehensive context present in longer texts, making it challenging to understand the intended meaning.
- **Multi-Topic.** Despite their brevity, short texts may encompass multiple topics with not enough elaboration on each topic.

Similar to short text, transcripts derived from video segments are brief and vague, lacking comprehensive context which pose significant challenges for traditional text classification algorithms (Lee and Dernoncourt, 2016; Zhou, 2017; Qiang et al., 2020). To characterize video segments based on text classification, it is necessary to investigate methods suitable for short text classification (STC). Next, we will review and analyze weakly supervised methods for STC.

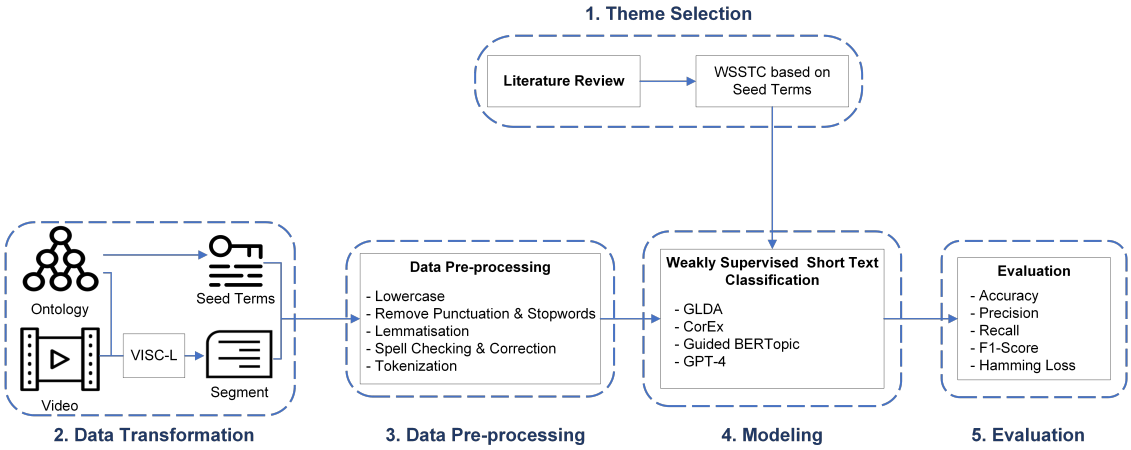

Figure 1: Framework of Weakly Supervised Short Text Classification for Characterising Video Segments. The performance of three WSSTC methods based on seed terms and GPT-4 is evaluated on transcripts of video segments in the health domain.

2.3 Weakly Supervised Short Text Classification

Weakly supervised learning is a comprehensive term encompassing numerous studies aimed at building predictive models through learning with different weak supervision types (Zhou, 2017):

- **Incomplete Supervision.** Only a small portion of data is labeled, and while the quantity of labeled data may be insufficient for effective model training, it is partially mitigated by a larger amount of unlabeled data that contributes to the process.
- **Inaccurate Supervision.** This refers to situations with potentially inaccurate labels, often observed in learning with label noise scenarios, where some training examples may be mislabeled.
- **Inexact Supervision.** This type of supervision provides only coarse-grained label information which may not be as precise as desired. Examples of inexact supervision include the utilization of knowledge bases, heuristic rules, and keywords.

Compared to supervised and unsupervised methods, weakly supervised learning addresses the need for a large amount of labeled data and often provides more interpretable classification results than unsupervised methods. Unlike deep learning methods, it doesn't require extensive training time and computational resources. Moreover, weakly supervised learning can handle attenuated data label signals, accommodate different label forms, and cater to various instances of inexact supervision. Consequently, several weakly supervised short text classification (WSSTC) approaches have been applied to STC.

Existing research in WSSTC has utilized knowledge bases (Türker et al., 2020), heuristic rules (Shu et al., 2020), and seed terms (Meng et al., 2019) as weak supervision signals to guide models to classify. Among these options, seed terms offer greater flexibility and simplicity. They do not require extensive time or specialized knowledge, only a basic understanding of each category. This significantly reduces the entry barrier for newcomers or when dealing with new domains. Therefore, in this project, we have chosen to adopt the seed terms method for WSSTC.

Guided Latent Dirichlet Allocation (GLDA)[1], Correlation Explanation (CorEx)[2], and Guided BERTopic[3] all allow for the provision of seed terms describing each category as weak supervision signals to enhance the text classification process. To understand the performance of these different methods, we will employ all three of the aforementioned methods.

2.4 Use of Large Language Models

The advancements in LLMs make them promising for applications in the education domain. For example,LLMs can provide learners with feedback on the educational content they create(Denny et al., 2022). Additionally, LLMs can generate corresponding solutions for exercises in teaching, opening up new possibilities and support for education(Savelka et al., 2023). However, existing research has primarily focused on exploring their generative capabilities, with limited studies on their STC abilities. Considering that LLMs are famous for their ability to provide extensive contextual information(Xie et al., 2021), this

[1]https://guidedlda.readthedocs.io/en/latest/
[2]https://ryanjgallagher.github.io/code/corex/overview
[3]https://maartengr.github.io/BERTopic/getting_started/guided/guided.html

research will also explore the STC capabilities of the currently most popular LLM, GPT4[1].

To compare the performance of different models and assess the STC capabilities of ChatGPT, this research will conduct a comparative analysis using GLDA, CorEx, Guided BERTopic, and GPT-4. The framework of this research, based on the selected techniques and methods, is shown in Figure 1. Next chapter will introduce our methodology.

3 METHODOLOGY

The experimental method is structured as follows: data sets and classification models.

3.1 Data Sets

We have implemented our methodology in the domain of Health Related Quality of Life Awareness (HRQLA). The **data** of our domain is the video segments with their transcript and characterisation (labels) that have been generated using Health Related Quality of Life Ontology[2] and (VISC-L). The transcript of the video segments will be used as an input to the selected models and the characterisation of these segments will be used as the ground truth to evaluate the models. Using VISC-L (Mohammed, 2022), 60 videos have been collected, segmented and characterised using the domain ontology and the result is: ENVIRONMENT (74), LEVEL OF INDEPENDENCE (246), PHYSICAL HEALTH(221), PSYCHOLOGICAL HEALTH (198), PERSONAL VALUES AND BELIEFS (2), and SOCIAL RELATIONSHIP (85) (Mohammed, 2024).

To normalize and standardize the transcripts and the seed terms, we employed Natural Language Processing (NLP) techniques for data pre-processing, which included lowercase, punctuation and stopwords removal, lemmatization, spell checking and correction, and tokenization. Then, we constructed the corresponding corpus, consisting of 2160 tokens. We partitioned our dataset into a training set (70%) and a test set (30%).

3.2 Classification Models

Data Vectorization. To transform video segment transcripts into a computable form for model input, we employed vectorization. GLDA utilizes a standard BOW approach for vectorization, representing transcripts as unordered bags of words, maintaining only word frequency and disregarding order. In contrast, CorEx works only on binary data, employing the Naïve Binarization approach based on binary BOW (Gallagher et al., 2017). Instead of retaining word frequencies, this method marks each word in the transcripts as either present (1) or absent (0). Guided BERTopic and GPT-4 employ embedding models to represent words as dense vectors in a high-dimensional space. This enables capturing semantic similarity through spatial proximity, ensuring comparable embeddings for words with similar contexts and encapsulating intricate relationships and nuances in their representations. Unlike Guided BERTopic, GPT-4 accepts raw text input and encodes it on its own. To align with other models that take vectorization of pre-processed data as input, we used both raw data and pre-processed data as inputs for GPT-4.

Model Implementation. The implementation procedures for all models and their execution in the STC tasks is illustrated in Figure 2.

GLDA is a variant of Latent Dirichlet Allocation (LDA) to discover latent or hidden topics within extensive text data (Jagarlamudi et al., 2012). It enhances LDA by incorporating seed terms, based on prior knowledge, to ensure the emergence of specific topics. CorEx, operating on information theory principles, aims to maximize mutual information (MI) between identified topics and observed data (Gallagher et al., 2017). Users can include domain knowledge using "anchor words." In our case, seed terms are used as anchor words, and CorEx maximizes MI $I(X;T)$ between words in transcripts and latent topics. Guided BERTopic employs BERT (Bidirectional Encoder Representations from Transformers) model to generate dense text embeddings, capturing semantic relationships (Grootendorst, 2022). It introduces user-defined seed terms to guide topic extraction towards specific areas. Our seed terms vocabulary is generated from the ontology, wherein the concepts of each theme are used to describe their respective topics. After training, we obtained our GLDA, CorEx, and Guided BERTopic models.

For GPT-4, we used the ChatGPT's Chat Completions API, which supports taking a list of messages as an input and returns the model-generated message as the output (OpenAI, 2023). It supports user-defined different roles in the conversation to guide GPT-4 in completing STC tasks. We use the "System" role to specify task requirements and instructions, while the "User" role is employed to submit transcripts that require classification. Finally, we parsed the informa-

[1]https://openai.com/research/gpt-4

[2]https://github.com/Health-Related-Quality-of-Life-Ontology/Ontology.git

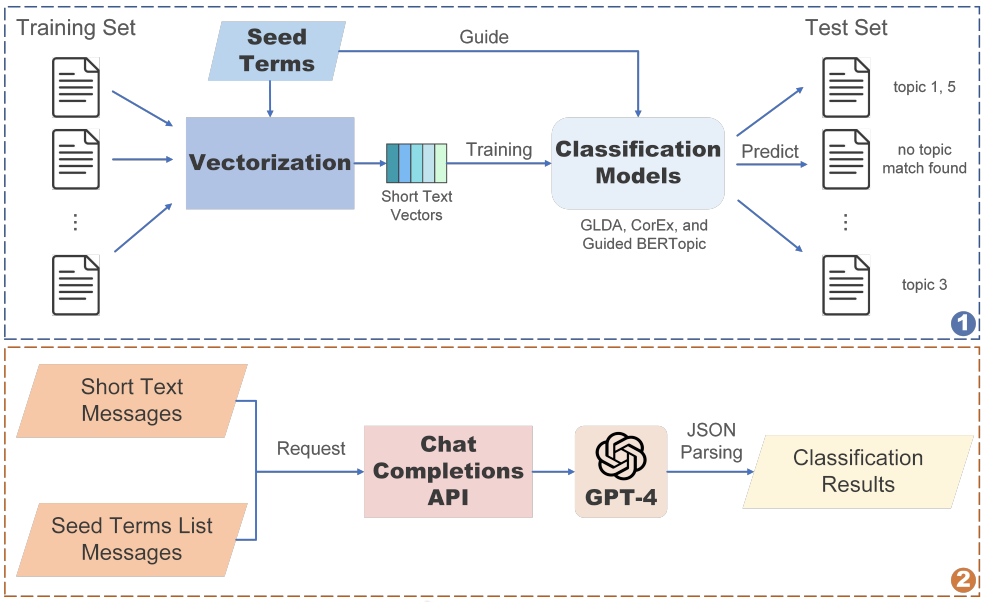

Figure 2: The diagram above shows the schematic representation of STC conducted by GLDA, CorEx, and Guided BERTopic under the guidance of seed terms (1). The diagram below illustrates the process of conducting STC with the assistance of seed terms by GPT-4 via Chat Completions API (2).

tion returned by GPT-4 in JSON format to obtain the classification results.

4 RESULTS

To evaluate the performance of different models, we employed the Hold-out method and considered Accuracy, Precision, Recall, F1 Score, and Hamming Loss. This section presents the comprehensive evaluation results of the five models on the test set based on the ground truth data. Additionally, it provides an in-depth analysis of the results by topics. Finally, it presents a comparative analysis between the classification results of the best-performing model and the ground truth data.

4.1 Models Performance

Based on the ground truth data, we evaluated the classification results of each model on the test set. The performance of all models on the evaluation metrics is presented in Table 1.

In this study, the CorEx model demonstrated superior performance in a specific STC challenge, achieving an accuracy of 62.12%. Its precision and recall rates, 89.20% and 78.45% respectively, emphasize its strength in accurate prediction and identifying true positives. A notably low HL of 5.57% further confirms its effectiveness. These results suggest that CorEx is well-suited to the characteristics of the data set used.

Conversely, both GLDA and Guided BERTopic models showed suboptimal results, with accuracies of 9.09% and 4.55%, respectively. A significant performance gap compared to CorEx is evident across all metrics, as highlighted by their higher HLs (39.27% for GLDA and 34.22% for Guided BERTopic), suggesting a greater likelihood of cross-label misclassifications. Despite their effectiveness in various NLP applications, their suitability for this specific data set warrants reevaluation.

The analysis of GPT-4's performance with both raw and pre-processed data reveals minimal differences, indicating that the text pre-processing stage did not significantly impact data quality. While GPT-4's recall is comparable to CorEx's (73.28% for raw data and 72.17% for pre-processed data against CorEx's 78.45%), it falls short in other metrics. This suggests that, despite GPT-4 can predict a certain amount of contextual information, its ability in STC tasks still requires further optimization.

4.2 In-depth Analysis by Topics

For a more comprehensive understanding, we conducted an analysis of the performance of the four algorithms on video segments across six distinct topics. The results of all evaluation metrics across topics are illustrated in Figure 3.

Table 1: Evaluation metrics of classification models on the test set against ground truth.

Models & Methods		Evaluation Metrics (%)				
		Accuracy	Precision	Recall	F1 Score	HL
WSSTC	GLDA	9.09	23.11	15.05	17.12	39.27
	CorEx	62.12	89.20	78.45	82.24	5.57
	Guided BERTopic	4.55	26.06	23.38	22.38	34.22
LLM	GPT-4 (Raw Data)	18.94	65.21	73.28	63.75	24.12
	GPT-4 (Pre-processed Data)	18.18	63.23	72.17	63.29	23.36

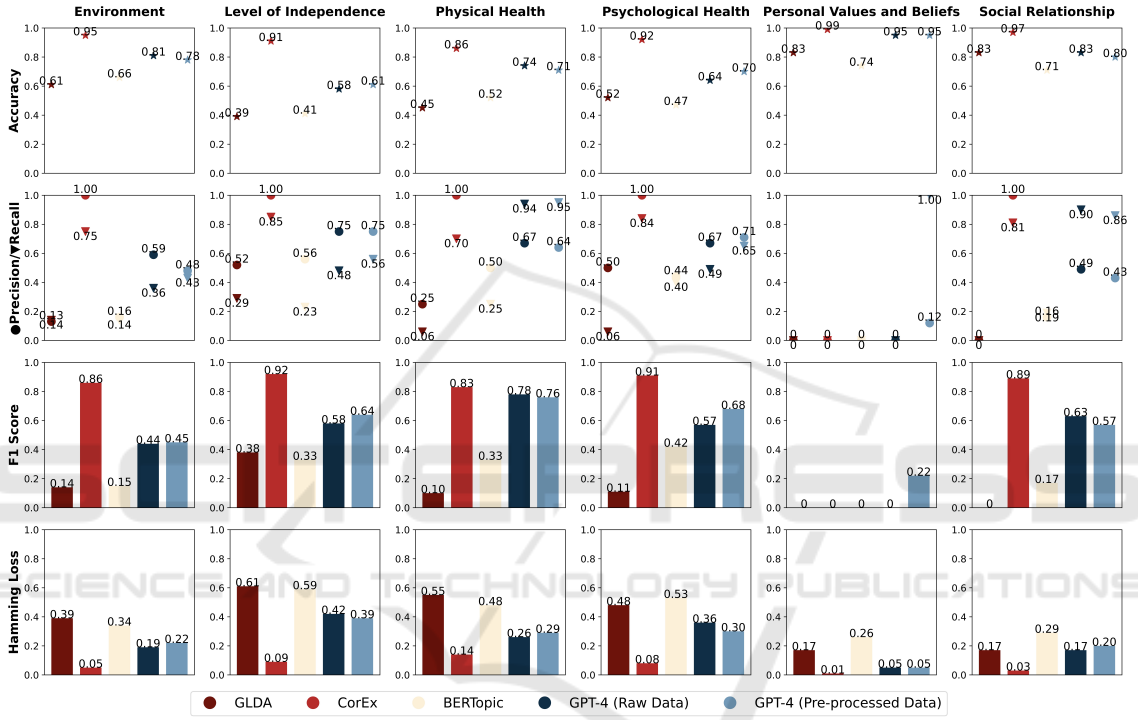

Figure 3: Evaluation metrics across topics of classification models on the test set against ground truth.

It's notable that in the "Personal Values and Beliefs" topic, metrics such as Precision, Recall, and F1 score are observed to be zero in some methods. This is attributed to the limited number of videos collected for this specific topic. CorEx emerged as a top performer, especially in the "Environment" topic, achieving a remarkable 0.95 accuracy and maintaining perfect precision in the first four categories. Its low HL in comparison to other models underlines its reliability. While GLDA and Guided BERTopic showed variable overall results, they excelled in specific topics. Specifically, in the "Social Relationship" topic, GLDA achieved an accuracy of 0.83, while Guided BERTopic achieved an accuracy of 0.71. However, the lower precision and recall in this category also indicate that there is potential for improvement in the overall performance of GLDA and Guided BERTopic.

GPT-4, when fed with raw data, displayed remarkable results in the "Physical Health" category, boasting a recall of 0.94. Additionally, its accuracy peaked at 0.95 for the "Personal Values and Beliefs" category. However, the precision metrics for this model require attention, especially for categories where it registered low values. In contrast, when fed pre-processed data, it exhibited strong balanced performance across categories. While its accuracy wasn't the highest, it didn't see significant drops across different topics. Some categories might exhibit higher precision (indicating fewer false positives), while others might have higher recall (indicating fewer missed actual positives). Notably, this model didn't display extreme values like 0 or 1 in various metrics. In sum, this approach seemed to have superior stability compared to other methods.

In in-depth analysis by topics, CorEx consis-

tently outshined others, particularly in the "Environment" category. While GLDA and Guided BERTopic demonstrated strong capabilities in certain topics, they still lag significantly behind other methods and exhibit inconsistent stability across different topics. The GPT-4 model, when leveraging raw data, manifested superior results in categories such as "Physical Health", yet its precision metrics across some topics leave room for refinement. Interestingly, the adaptation of pre-processing techniques with GPT-4 showcased a more balanced performance.

4.3 CorEx & Ground Truth Data

Based on the above analysis, CorEx emerged as the best-performing model among the four. To understand the discrepancies between its classification results and the ground truth data, we conducted a comparison. According to our analysis, out of the 132 samples in the test set, CorEx correctly classified 83 samples, which accounts for 62.88% of the total samples. The discrepancies can be categorized into two types: 14 samples that couldn't be classified at all (fully under-classification) and 35 samples that were under-classified.

To understand the reasons behind these discrepancies, we analyzed the MI during the classification process for these 49 samples. The results indicate that CorEx struggles to correctly classify some concepts with high MI, such as "feeling" (3.7350) and "information" (3.1183). This difficulty may be attributed to its inherent limitations. Since CorEx relies on a greedy algorithm, it may not always identify the global optimal solution, settling for a local optimum instead. Consequently, terms with high MI to certain topics might not be assigned to them. Additionally, the BOW vectorization approach, when applied to short texts, can result in sparse representations, which are further affected by noise such as paralinguistic elements. Paralinguistic elements are non-verbal features that accompany speech and contribute to communication, such as "um" and "uh." This noise can inadvertently influence the model's decision-making process. Certain key concepts in CorEx displayed near-zero MI, , such as "disease" (0.0032) and "breath" (0.0001), leading to miss-classification. This could be attributed to the presence of noise within the dataset, which might depress the MI values of essential concepts, causing them to be overlooked by the model.

Despite the mentioned limitations, the performance of CorEx in this task is commendable. It does not rely on deep learning methods or LLMs. Instead, it achieves good performance in this task just with the guidance of seed terms as weak supervision signals.

5 DISCUSSION & CONCLUSION

In this study, we proposed guiding the model for classifying transcripts of video segments on HRQLA domain by using the concepts of each topic from ontology as weak supervision signals. We compared the performance of different models, including GLDA, CorEx, Guided BERTopic, and GPT-4. The main challenges were limited contextual information provided by transcripts of video segments, given that they are considered short text. Additionally, there is a limitation in the number of samples available for training. Some topics have fewer videos, resulting in a scarcity of samples for those particular topics.

Our analysis revealed that among the four models, CorEx performed the best. This is attributed to CorEx leveraging MI optimization to uncover hidden patterns in the texts, making it proficient at capturing intricate relationships among various topic concepts present in the transcripts. Although CorEx's algorithm is based on a greedy approach and may occasionally suffer from miss-classification due to local optima, achieving such results with just the guidance of seed terms is commendable.

However, the performance of GLDA and Guided BERTopic was less satisfactory. GLDA, despite utilizing seed terms as guiding words, often deviated significantly from the original topics. The adverse impact of irrelevant noise in the dataset could be a contributing factor to this deviation. As for Guided BERTopic, it relies on cosine similarity measurements between all seed words for each topic and the recorded transcripts. It neglects cases where only individual concepts or subsets of concepts appear, leading to suboptimal performance.

We evaluated GPT-4 using both raw text and pre-processed text as inputs, and interestingly, the results showed minimal differences between the two modes. This suggests that we did not lose crucial information during the data pre-processing stage. While GPT-4's consistent performance is commendable, it still lags behind CorEx in this task. Despite GPT-4's popularity and excellent performance in many NLP tasks, there is room for improvement in its STC capabilities.

This innovation provides a powerful tool for educators and learners to efficiently learn from videos, aligning specific video segments with relevant learning domains and objectives. As a result, educators can tailor instructional materials more precisely to meet diverse learning needs, enhancing the overall educational experience. Learners can access video content

closely aligned with their learning interests and requirements, thereby increasing their engagement and comprehension. In summary, this research not only advances the field of STC but also offers a practical solution to enhance video-based learning.

Future work can consider increasing the size of text which includes: increasing the number of video segments and the duration of each segment to include more transcript text lines. This to allow us observe how it affects the performance of various models within the given short text size constraints. Once the video segments are characterised, we have developed a framework to link video segments to support learning of specific domain concepts(Mohammed, 2024).

We envisage applications of the work in other domains related to using videos for life-wide learning based on other people's experiences, e.g. communication, project management, empathy, where automated video characterization can enable efficient video linking to support informal learning.

REFERENCES

Barrett, M. D. (2018). *Reference framework of competences for democratic culture*. Council of Europe.

Bywater, J. P., Floryan, M., and Chiu, J. L. (2021). Discs: A new sequence segmentation method for open-ended learning environments. *Springer*, pages 88–100.

Commission of the European Communities (2000). *A memorandum on lifelong learning*. European Commission.

Denny, P., Sarsa, S., Hellas, A., and Leinonen, J. (2022). Robosourcing educational resources – leveraging large language models for learnersourcing. *arXiv*.

Dodson, S., Roll, I., Harandi, N. M., Fels, S., and Yoon, D. (2019). Weaving together media, technologies and people. *Inf. and Learning Sciences*, 120:519–540.

Gallagher, R. J., Reing, K., Kale, D. C., and Steeg, G. V. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the ACL*, 5:529–542.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv*.

Jagarlamudi, J., Daumé, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. *Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems*, 36:1–30.

Luca, C., Canale, L., and Farinetti, L. (2019). Visa: A supervised approach to indexing video lectures with semantic annotations. *2019 IEEE 43rd Annual Computer Software and Applications Conference*.

Meng, Y., Shen, J., Zhang, C., and Han, J. (2019). Weakly-supervised hierarchical text classification. *Proceedings of the 33rd AAAI Conference*, 33:6826–6833.

Mitrovic, A., Dimitrova, V., and Weerasinghe, A. (2016). Reflective experiential learning: Using active video watching for soft skills training. *Proceedings of the 24th Int. Conference on Computers in Education*.

Mohammed, A. (2022). Generating narratives of video segments to support learning. In *23rd International Conference on Artificial Intelligence in Education*, pages 22–28. Springer.

Mohammed, A. (2024). *Generating Video Narratives to Support Learning*. PhD thesis, University of Leeds.

OpenAI (2023). Chat completions api. https://platform.openai.com/docs/guides/text-generation/chat-completions-api [Accessed: (June 2023)].

Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Redeker, C., Leis, M., Leendertse, M., Punie, Y., Gijsbers, G., Kirschner, P. A., Stoyanov, S., and Hoogveld, B. (2012). The future of learning: Preparing for change.

Sablić, M., Mirosavljević, A., and Škugor, A. (2020). Video-based learning (vbl)—past, present and future: an overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning*.

Savelka, J., Agarwal, A., Bogart, C., and Sakr, M. (2023). Large language models (gpt) struggle to answer multiple-choice questions about code. *arXiv*.

Schlotterbeck, D., Uribe, P., Jiménez, A., Araya, R., , v., and Caballero, D. (2021). Tarta: Teacher activity recognizer from transcriptions and audio. *Lecture Notes in Computer Science*, pages 369–380.

Shu, K., Mukherjee, S., Zheng, G., Awadallah, A. H., Shokouhi, M., and Dumais, S. T. (2020). Learning with weak supervision for email intent detection. *arXiv*.

Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. (2014). Short text classification: A survey. *Journal of Multimedia*, 9.

Tseng, S.-S. (2021). The influence of teacher annotations on student learning engagement and video watching behaviors. *International Journal of Educational Technology in Higher Education*, 18.

Türker, R., Zhang, L., Alam, M., and Sack, H. (2020). Weakly supervised short text categorization using world knowledge. *Springer eBooks*, pages 584–600.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. *arXiv*.

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5:44–53.