5. Toward the Automatic Identification of Isotopies

Alice Fedotova and Alberto Barrón-Cedeño

ABSTRACT

The rise in processing power, combined with advancements in machine learning, has resulted in an increase in the use of computational methods for automated content analysis. Although human coding is more effective for handling complex variables at the core of media studies, audiovisual content is often understudied because analyzing it is difficult and time-consuming. The present work sets out to address this issue by experimenting with unimodal and multimodal transformer-based models in an attempt to automatically classify segments from the popular medical TV drama *Grey's Anatomy* (ABC, 2005-) into three isotopies that are typical of the medical drama genre. To approach the task, this study explores two different classification approaches: the first approach is to employ a single multiclass classifier, while the second involves using the one-vs-the-rest approach to decompose the multiclass task with a series of binary classifiers. We investigate both these approaches in unimodal and multimodal settings, with the aim of identifying the most effective combination of the two. The results of the experiments can be considered promising, as the multiclass multimodal approach results in an F1 score of 0.723, a noticeable improvement over the F1 of 0.686 obtained by the one-vs-the-rest unimodal approach based on text.

KEYWORDS

Multimodal content analysis; deep learning; transformers; multimodality; medical dramas.

Introduction

In the field of media studies, content analysis is an established methodology for the study of audiovisual products. A central aspect of content analysis is *coding*, which consists in assigning units of analysis to categories for the purpose of describing and quantifying phenomena of interest (Krippendorff 1980: 84-5). Previous research has identified three fundamental categories or "isotopies" that characterize the medical drama genre: the professional plot, the sentimental plot, and the medical cases plot. In the context of medical dramas, content analysis can be conducted by assigning isotopies to segments, i.e. portions of video "characterized both by space–time–action continuity and invariance in the thematic-narrative elements" (Rocchi and Pescatore 2022: 3). This poses a challenge for automated approaches, as modern segmentation algorithms are not effective in identifying units that are relevant for the three isotopies. Additionally, coding requires trained annotators with a significant degree of expert knowledge and a good understanding of content analysis. Recognizing the complexity of the task and the need for more effective strategies, we experiment with unimodal and multimodal transformer-based models to evaluate the possibility of streamlining the content analysis process for medical dramas. With this objective, we formulate the following research questions:

- **Research Question 1:** Is it better to approach the task with a single multiclass model or a one-vs-the-rest approach?
- **Research Question 2:** Which modality is more informative for the task of predicting the isotopies?
- **Research Question 3:** Does the inclusion of keyframes in addition to the subtitles result in higher performance as compared to only using the subtitles?

To answer our research questions, we first create a multimodal corpus by combining subtitles and keyframes extracted from 17 seasons of *Grey's*

Anatomy (ABC, 2005-), one of the longest-running medical drama series. Three deep learning models, namely CLIP, BERT, and MMBT, are trained using this corpus to explore the impact of different modalities on the identification of the isotopies. Additionally, we investigate two different approaches to the classification problem: a multiclass approach, which considers all isotopies simultaneously, and a one-vs-the-rest approach, which identifies one isotopy at the time.[1] The results of the experiments are promising, with the multiclass multimodal approach obtaining an F1 score of 0.723. Furthermore, our findings suggest a relationship between the two approaches and the modalities involved, as well as differences in terms of the contribution of the individual modalities.

Related Work

With its ability to engage several human faculties at once, audiovisual content can convey information in a more multifaceted way compared to static images or text. However, the addition of the time element through shots and scenes makes the task of understanding the content of a video complex. One of the biggest challenges in the fields of natural language processing and computer vision is developing the ability for machines to analyze and summarize audiovisual products, making them more searchable and accessible (Tapaswi 2016: 3). As multimodal data often represents an object from different viewpoints, which can be complementary in contents, it can potentially be more informative than unimodal data. However, there are also instances where the modalities end up competing with each other, causing multimodal models to underperform compared to the unimodal ones (Huang et al. 2021: 10944).

Compared to visual and auditory information, textual features are less explored for video understanding (Weng et al. 2021: 4843). In the broader context of movies and TV shows, speech may sometimes be correlated with the action (e.g., "Raise your glasses to..."), but it is more frequent for it to be completely uncorrelated (Nagrani et al. 2020: 10318). In the field of sentiment analysis, related work has been conducted on the TV show *Friends* (NBC, 1994-2004). Zahiri and Choi (2017: 6-7) employ a CNN

[1] The scripts used for the experiments are available at https://github.com/TinfFoil/isotopy-identification/.

architecture with word2vec embeddings for the purpose of detecting emotions from written dialogue, obtaining accuracies of 37.9% and 54% for fine- and coarse-grained emotions respectively. They observe that emotions are not necessarily conveyed in the text, and that disfluencies, metaphors, and humor make the task particularly challenging.

Vision-and-language approaches to video understanding can be divided into two types: one based on using a single frame, and another based on extracting multiple frames (Sun et al. 2019, Zhu and Yang 2020). In the context of video, the second approach is more common, as it is reasonable to assume that training an effective video-and-language model requires lots of samples from the video channel (Lei et al. 2022: 11). As demonstrated by Li et al. (2021: 7), leveraging both video and subtitles achieves the best performance on the VALUE benchmark (Li et al. 2021), which includes 11 video understanding tasks from a variety of datasets and video genres. A similar result is reported by Liu et al. (2020: 10906) on the task of video-and-language inference, which consists in analyzing a video clip paired with a natural language hypothesis and determining whether the hypothesis is supported or contradicted by the information conveyed in the video.

However, it is actually an open question whether training a model using multiple frames is beneficial for downstream tasks, and if so, whether the gains in performance justify the significant increase in computational costs (Lei et al. 2022: 11). Despite the fact that most video-and-language models are typically trained using multiple video frames, some studies suggest that strong performance on challenging benchmarks can be achieved using just a single frame (Lei et al. 2022, Buch et al. 2022). Furthermore, the difficulty of making recognition decisions is intrinsically linked to the type of category being classified. For instance, recognizing static subjects like dogs and cats, or sceneries such as forests or seas, may only require a single frame. However, distinguishing more complex actions, such as "walking" versus "running", often requires more frames (Wu et al. 2019: 1284).

To the best of our knowledge, this is the first work on narrative classification for the medical drama genre. In the context of cinema, a similar work is the Movie Narrative Dataset (MND), introduced by Liu et al. (2023). MND consists of 6,448 annotated scenes from 45 movies, manually labeled by multiple annotators into 15 key story elements. To benchmark the task of classifying scenes based on their narrative function, the authors of MND utilized an XGBoost classifier trained on temporal features and character co-occurrence patterns. The classifier obtained an F1 score of 0.31, which, while still leaving room for improvement, is statistically significant and outperforms a static baseline classifier. Unlike Liu et al. (2023), we adopt a single-frame vision-and-language approach. This choice is motivated by the previously mentioned studies showing the potential of using only a single frame (Lei et al. 2022, Buch et al. 2022). Additionally, the decision to consider a single frame is influenced by the substantial increase in computational costs associated with analyzing multiple frames (Lei et al. 2022: 11), which presents significant challenges in terms of resource requirements and processing time.

Dataset

The present work builds upon the Medical Dramas Dataset introduced by Rocchi and Pescatore (2022: 2-3). For the purpose of the experiments, we extracted 17 seasons of annotated data from the TV show *Grey's Anatomy* (2005-), for a total of 367 episodes and 244 hours of video.[2] Isotopy assignment, also referred to as 'coding', was conducted according to a three-step content analysis protocol. First, three isotopies underlying the medical drama genre were identified: the medical cases plot, the professional plot, and the sentimental plot. According to Pescatore and Rocchi (2019: 111-112), the isotopies can be defined as follows:

The medical cases plot (MC) is related to the storylines that usually change between each episode, introducing new narrative elements and a variety of characters into the hospital setting.

The professional plot (PP) deals with the relationships and dynamics within the hospital among doctors and other medical staff.

The sentimental plot (SP) comprises the emotional and personal relationships between the main characters throughout the series. It covers a wide sphere of emotions such as friendship, love, empathy, and conflict.

The second step involved breaking down each episode into segments. For each segment, start and end times were marked. This aspect is especially important, as it allowed the subsequent alignment with the text of the subtitles. The third phase, i.e. the actual coding phase, followed the identification of the segments. During this step, the appropriate isotopies were assigned to

[2] The data was mainly obtained from https://doi.org/10.17605/OSF.IO/24TUS, with the addition of 5 seasons of unpublished data provided by the authors.

each previously identified segment, taking into account their development over time, and not treating them as independent segments. A weight from 0 to 6 was assigned to each of the plots. If a segment could only be attributed to a single plot, a weight of 6 was assigned to that plot and a weight of 0 to the other two. When there were overlaps between narrative lines, a weight was assigned to each of the co-occurring narratives according to their relevance in the segment. In some cases, segments were not attributable to either of the isotopies and all three were marked as "NA" (Rocchi and Pescatore 2022: 3).

Data Extraction

The availability of start times and end times for each segment allowed for the alignment of the dataset with another source of data tagged with temporal information: the subtitle track of the episodes. Each subtitle has four parts in a SubRip Subtitle (SRT) file:[3] a counter indicating the number of the subtitle; start and end timestamps; one or more lines of text; and an empty line indicating the end of the subtitle. By relying on these features, the SRT files were processed to extract the timestamps and the text of the subtitles.

For the purpose of aligning the subtitles with the data obtained from the Medical Dramas Dataset, a method for assigning each of the subtitles to the corresponding segment was then identified. Inspired by Tapaswi et al. (2015: 5), in which subtitles appearing at video shot boundaries were attributed to the shot which has a majority portion of the subtitle, the mean of each subtitle's timespan was used as the criterion for the alignment. For example, given a subtitle that starts at 00:00:00.804 and ends at 00:00:02.701, the mean is 00:00:01.752. If a segment starts at 00:00:00.000 and ends at 00:00:07.000, then the subtitle is part of that segment. By doing so, a subtitle that overlaps with two different segments is assigned to the one where it appears on the screen for the longest amount of time.

In addition to aligning the subtitles, keyframes were also extracted from each of the episodes. A script based on OpenCV (Bradski 2000), an open-source computer vision library, was developed to accomplish this task. For each video, the midpoint of each segment was calculated based on the start and end times of the segment. The corresponding keyframe is then extracted

[3] https://docs.fileformat.com/video/srt/ (last accessed 16-07-2023).

and stored as a JPG file. Table 1 illustrates a few examples from the corpus, consisting of different segments and their timestamps, as well as the text obtained from the subtitles, the filenames of the keyframes and the assigned isotopies. Segments up to 00:00:49 are missing in this case because they were labeled as NA. An example of a keyframe is also shown in Figure 1.

id	segm_start	segm_end	pp	sp	mc	img_name
S13E01_0	00:00:49	00:02:18	0	6	0	S13E01_0.jpg
Meredith: Don't you wish you could just take it back... That thing you said, that thing you did. […] We can't undo the past. `Cause the future keeps coming at us.						
id	segm_start	segm_end	pp	sp	mc	img_name
S13E01_1	00:02:18	00:02:36	0	2	4	S13E01_1.jpg
[Siren wails] Isaac: What do we got? We got a male, mid 20s. […] We'll need a CT. All right, let's get him to Trauma One. Let's go. Page Avery!						
id	segm_start	segm_end	pp	sp	mc	img_name
S13E01_2	00:02:36	00:03:18	0	6	0	S13E01_2.jpg
Two champagnes. You got it. I thought you were dancing with Maggie. […] Take a breath. What happened to DeLuca?						

Table 1
Some instances from the resulting corpus. The text obtained from the subtitles has been shortened for displaying purposes.

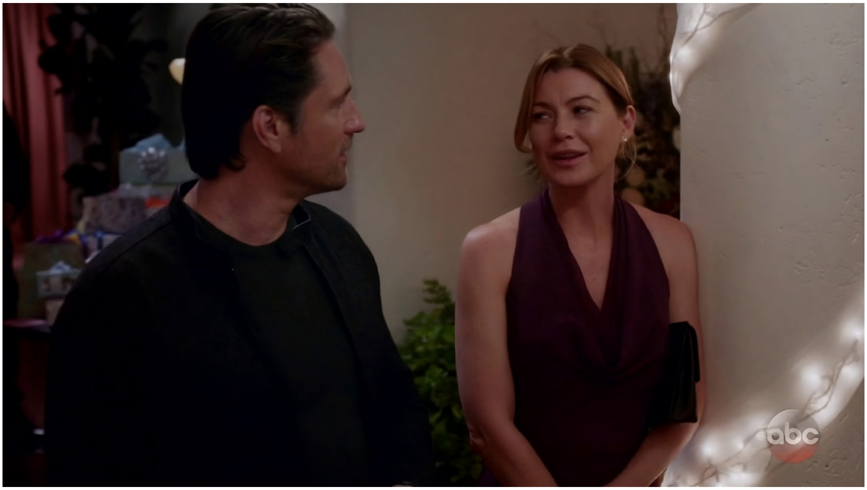

Figure 1
Keyframe 13x01.

Data Preprocessing and Description

The preprocessing of the corpus involved several steps designed to refine and improve the quality of the data and was mainly conducted using the NLTK library (Bird et al. 2009). Most importantly, segments containing nine subtitles or less in which stopwords and consecutive repeated words constituted more than 65% of the total tokens, were removed. In addition to this, other preprocessing steps included removing song lyrics (e.g., "♪ *I don't want to wait...*♪"); song names (e.g., "*[Lorde's 'Team' playing]*"); subtitle author's names (e.g., "*Telescript by Raceman, Subtitles/Sync by Bemused*"); italics tags (e.g., "<i>" and "</i>" in "*I'm <i>really</i> sorry*"); hesitations (e.g., "-he" in "*He-he doesn't... He doesn't mean that*"); hyphens indicating dialogue between different characters (e.g., "*-Is he talking? -Yeah.*"); and segments containing only sounds (e.g., "*[Whistles]*").

Labels were also preprocessed as part of the data preparation, with the original range of [0, 6] discretized into binary values of {0, 1}. Values in the interval [0, 2] were mapped to 0 and values in the interval [4, 6] were mapped to 1; as a result, segments with label combinations 330, 303, and 033 were removed as they could not be discretized into the required binary representation. This is because segments having combinations such as PP=3, SP=3, MC=0 are characterized by two different isotopies having the same weight. The counts of the instances per class before and after discretization are illustrated in Table 2. Although some of the granularity in the original data is lost, the main advantage of this approach is that it simplifies the classification task by reducing the number of classes, which enables the model to focus on identifying those segments where there is a complete or mostly complete correspondence to one of the isotopies.

Before discretization				After discretization			
Values	PP	SP	MC	Values	PP	SP	MC
0 1 2	13,668 321 667	8,718 310 751	11,641 245 621	0	13,690	8,907	11,381
4 5 6	368 156 2,775	445 297 7,340	394 233 4,981	1	3,299	8,082	5,608

Table 2
Corpus label distribution before and after discretization.

The resulting corpus contains 276,357 subtitles, which are grouped into 16,989 labeled segments. The corpus has a total of 2,260,655 tokens (38,629 types) and the mean length of a subtitle is 8.430 ± 3.921 tokens. Each segment consists of 1 to 74 subtitles, and about 95.7% of the segments (16,272) contains up to 37 subtitles. The dataset is imbalanced, with the sentimental plot class being the most represented out of the three (8,082 positive instances). The least represented class is the professional plot, with 3,299 positive instances, while the medical cases class has a total of 5,608 positive instances.

Experiments

To address our research questions, we explore two different classification approaches to determine which one is better suited for the problem at hand: the first approach is to employ a single multiclass classifier, while the second involves using the one-vs-the-rest approach. Although neural networks can handle the multiclass problem by directly predicting one of the three possible target classes, using the one-vs-the-rest (OvR) strategy may be beneficial in certain situations. This approach, also known as one-vs-all (OvA), consists in decomposing the task into *n* binary classifiers, each trained to distinguish between one class and the rest. The final prediction is made by selecting the class associated with the classifier that outputs the highest probability (Aly 2005: 1-4). We investigate the multiclass and one-vs-the-rest approaches for both unimodal and multimodal settings. For the multiclass approach, we first fine-tune and evaluate a unimodal textual and a unimodal visual model, and then a multimodal one. For the one-vs-the-rest approach, we do the same for each unimodal binary sub-problem, and then repeat the problem decomposition approach in the multimodal setting as well.

For the unimodal textual setting, we use the bert-base-uncased implementation of BERT from the HuggingFace library (Devlin et al. 2018). The model is fine-tuned exploring epochs ∈ [1, 2, 3] with a batch size of 16, one of the batch sizes recommended by the authors of BERT (Devlin et al. 2018). For optimization, we employ the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of 1e-5 and an epsilon value of 1e-8. We encode the training, validation, and test datasets with BertTokenizer[4]

[4] https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#berttokenizer (last accessed 16-07-2023).

and pad the sequences to a maximum length of 512. To adapt the unimodal model for the two approaches, we modify the num_labels parameter of BERT, setting it to 3 for multiclass classification and 1 for binary classification. For multiclass, we use the default Cross Entropy loss function that is computed by BERT when num_labels > 1 (Devlin et al. 2018). For binary classification in the one-vs-the-rest scenario, we use the Binary Cross Entropy with Logits loss from PyTorch.[5]

For the multimodal setting, we use the Multimodal Bitransformer (MMBT) model. Introduced by Kiela et al. (2019), MMBT incorporates the strengths of the transformer architecture and adapts it for processing both textual and visual inputs. To further enhance the capabilities of MMBT, we follow Muti et al. (2022) in using OpenAI's CLIP (Radford et al. 2021) as the visual encoder instead of the default ResNet-152 architecture used by MMBT. As for preprocessing, we use the Pillow library (Clark 2015) to prepare 288x288 pixel versions of all frames by rescaling and padding, while also maintaining the original aspect ratio of the frames (Neskorozhenyi 2021). We then slice the frames into three equal parts to obtain four vectors: a vector for each of the parts that encode spatial information and one for the whole frame. The visual feature extractor of CLIP is RN50x4, a modified version of ResNet-50 which has been shown to be particularly effective for vision-and-language tasks (Shen et al. 2021: 5-8).

As for the textual encoder, we again use bert-base-uncased so as to be able to compare the performance of MMBT and BERT. We fine-tune the MMBT architecture by exploring epochs ∈ [1, 2, 3] with a batch size of 8 and a gradient accumulation of 20 steps to reduce memory usage. For optimization, we employ the MADGRAD optimizer (Defazio and Jelassi 2022) with a learning rate of 2e-4. As for BERT, we adhere to the preprocessing and parameters used in the unimodal textual setting. Given that MMBT is largely based on BERT's architecture, the num_labels parameter and the loss functions are also configured in the same way as BERT. We maintain these choices for the unimodal visual model based on CLIP, with the exception that we do not use the textual encoder. As for the CLIP-based model, we leave RN50x4 as the feature extractor and we follow Wei et al. (2022) in using a batch size of 16.

[5] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss (last accessed 16-07-2023).

Models are evaluated using the F1-measure, a metric combining precision, which measures the accuracy of positive predictions, and recall, which measures the model's ability to identify all positive instances (Géron 2017: 84-87). Macro-averaging the F1 scores for each class gives equal weight to each class regardless of the number of instances. The models which obtained the most promising results during 10-fold cross-validation are illustrated in Table 3. A final evaluation of these models is then carried out on an independent test set that was not used during training to assess the models' performance on unseen data: in the case of the multiclass approach, the model which obtained the highest macro-averaged F1-measure on the validation set is evaluated on the test set; in the case of the one-vs-the-rest approach, the best-performing individual binary classifiers based on validation F1 score are ensembled to obtain the final prediction on the test set. Table 4 reports the results of this final evaluation.

		Multiclass	One-vs-the-rest						
Model	**(e)**	**All**	**(e)**	**PP**	**(e)**	**SP**	**(e)**	**MC**	**All**
CLIP	(3)	0.553	(3)	0.444	(2)	0.710	(3)	0.593	0.582
BERT	(3)	0.716	(2)	0.619	(2)	0.815	(2)	0.711	0.712
MMBT	(3)	0.736	(3)	0.580	(3)	0.825	(3)	0.741	0.715

Table 3
Validation F1 scores of the best models. (e) refers to the number of epochs.

		Multiclass	One-vs-the-rest						
Model	**(e)**	**All**	**(e)**	**PP**	**(e)**	**SP**	**(e)**	**MC**	**All**
CLIP	(3)	0.536	(3)	0.443	(2)	0.696	(3)	0.559	0.566
BERT	(3)	0.672	(2)	0.563	(2)	0.788	(2)	0.706	0.686
MMBT	(3)	0.723	(3)	0.592	(3)	0.818	(3)	0.728	0.713

Table 4
Test F1 scores of the best models. (e) refers to the number of epochs.

Discussion

As for RQ1, the answer is not straightforward. The approach which resulted in the best-performing model is the direct multiclass approach. Specifically, multiclass MMBT trained over 3 epochs achieved the highest macro-averaged F1 score on the test set: 0.723. This result could be attributed to the ability of the multiclass MMBT approach to better handle correlations between different classes, a feature not captured by the one-vs-the-rest approach, which treats each class independently. It is possible that the added visual information allows MMBT to disambiguate instances more effectively than the multiclass BERT model, resulting in one-vs-the-rest being more effective for BERT: F1 0.686 for one-vs-the-rest compared to 0.672 for multiclass. CLIP also benefits from the one-vs-the-rest approach. This could be due to the fact that one-vs-the-rest is more suitable for unimodal models, as a similar trend also arises when it comes to multiclass BERT compared to one-vs-the-rest BERT.

However, it should be noted that the one-vs-the-rest approach results in a noticeable increase in computational cost compared to training a single multiclass model. On two NVIDIA Quadro P4000 8GB GPUs, 10-fold cross-validation with MMBT required 8 hours for the multiclass model and 24 hours for the three binary models in the one-vs-the-rest approach. This is in part due to the smaller batch size of 8, which was chosen due to hardware limitations, although the difference is also noticeable in the case of BERT, which required 5 hours for multiclass and 16 for one-vs-the-rest with a batch size of 16. Approaches based on CLIP took about the same time. Furthermore, when considering BERT, the difference between multiclass and one-vs-the-rest is fairly small on the validation set, although one-vs-the-rest performed noticeably better on the test set. As the best approach to the task depends on both the modalities involved and the computational resources that are available, we will answer RQ2 and RQ3 by considering both settings.

In order to address RQ2, we compare the results obtained by the two unimodal approaches to determine which modality is more informative for the task of predicting the isotopies. On the test set, one-vs-the-rest BERT achieved an F1 score of 0.686, while one-vs-the-rest CLIP obtained a significantly lower F1 score of 0.566 (cf. Table 4). The difference between the two modalities is also evident in the multiclass setting, where BERT obtained an F1 score of 0.672 compared to CLIP's 0.536. As for RQ2, we can conclude that BERT performs better than CLIP, which suggests that the text might be more informative than the keyframes for the task of predicting the isotopies. It should be noted, however, that the models based on CLIP were limited by the fact that only a single keyframe was considered for each segment. Given the average length of the texts available to BERT, it is clear that the textual models not only had access to more information but could also analyze dialogue at different points in time, unlike CLIP which looks exclusively at the midframe of a segment. To overcome this limitation, an approach that takes into consideration multiple frames or a more systematically-chosen single frame could be developed.

Moving on to RQ3, we proceed to assess whether the combination of keyframes and subtitles resulted in higher performance by comparing the F1 scores of MMBT and BERT. As shown in Table 4, the best-performing MMBT model, i.e. multiclass MMBT, obtained an F1 score of 0.723, which is noticeably higher than multiclass BERT's F1 score of 0.672. The same is true in the one-vs-the-rest setting. Although not as effective for MMBT as it was for BERT, one-vs-the rest MMBT resulted in an F1 score of 0.713, which is still significantly better than one-vs-the-rest BERT's improved F1 of 0.686. Overall, multiclass MMBT's F1 score of 0.723 is the highest across all models and configurations. Considering RQ3, we can conclude that using a multimodal approach can result in a noticeable improvement over the text-only BERT model. Given the limitations presented in RQ2, we can consider this result to be promising, as it suggests that integrating more information from the visual channel can improve the performance of the model regardless of the approach that is being used, although the improvement is noticeably more pronounced in the multiclass setting compared to one-vs-the-rest.

In summary, one-vs-the-rest appears to be more effective for unimodal models, while textual features proved to be more informative than keyframes for predicting the isotopies. Overall, the improvement obtained by MMBT over BERT shows that the information from the visual channel complements the one that is contained in the dialogues. However, although one-vs-the-rest CLIP and BERT resulted in better generalization, the problem decomposition approaches would still require a longer training time compared to the best-performing model-approach combination, multiclass MMBT. Hence, addressing the task using a single multiclass MMBT model would still be recommended over one-vs-the-rest BERT. The second-best option would be to train a multiclass BERT model, which would be less computationally expensive but also less effective than MMBT. Regardless of the approach, exploring more parameters, such as different batch sizes or learning rates, would be the most immediate next step.

In the broader context, automated content analysis for isotopy identification, a domain which has been previously unexplored, can greatly benefit from multimodal approaches. Despite some limitations, these results also indicate the potential of single-frame approaches for the task of multimodal video classification.

Conclusions

This study examined three research questions to evaluate various methods for automatic isotopy identification in the context of TV medical dramas. The first research question focused on comparing, for all models, the performance of a direct multiclass approach versus a one-vs-the-rest approach. The second research question aimed to determine the most informative modality for the classification task. The third research question involved investigating whether the inclusion of keyframes in addition to subtitles resulted in better performance compared to just using the subtitles. In order to answer these research questions, we created a multimodal corpus by expanding on the Medical Dramas Dataset introduced in Rocchi and Pescatore (2022: 2-3). 17 seasons of annotated data were extracted from the TV show *Grey's Anatomy* (2005-), for a total of 367 episodes and 244 hours of video. Textual features were extracted by temporally aligning the subtitles with the segments, while visual features were obtained by extracting a frame, referred to as a keyframe.

The findings from this work are promising, indicating that it is indeed possible to leverage deep learning models to automatically identify the distinctive isotopies of the medical drama genre in the context of *Grey's Anatomy* (2005-). We observed that the multimodal MMBT model performed significantly better compared to the text-only BERT model and the image-only CLIP model. More specifically, MMBT achieved the top F1 score of 0.723, compared to BERT's highest F1 score of 0.686, thus shedding light on the potential benefit of incorporating visual information alongside textual data. We have also examined different approaches to the problem, observing that the one-vs-the-rest approach appears to be more beneficial in the case of unimodal models. It is possible that the added visual information allows MMBT to disambiguate instances more effectively than multiclass BERT, which could explain why this is the only setting in which multiclass worked better than one-vs-rest. The textual information proved to be more informative than the visual data, highlighting the importance of dialogue for isotopy identification.

The potential for future work is vast, as there are many aspects that could be further improved to enhance the performance of the models. For example, future research could delve into a more systematic methodology for frame selection, which in this study was limited to only the midframes of the segments. Apart from investigating more systematic approaches for frame selection, additional avenues for further research might include the adoption of a dual-stream model as an alternative to the single-stream architecture of MMBT. Existing research suggests that dual-stream models can obtain better results thanks to their co-attention mechanism, which enables them to handle complex relationships between the modalities (Du et al. 2022: 5437). Moreover, cross-lingual transfer could be explored by experimenting with multilingual transformer-based models like mBERT (Devlin et al. 2018) or XLM-RoBERTa (Conneau et al. 2020). Given the availability of subtitles in other languages, this approach could also lead to improvements and open up the possibility of analyzing other shows pertaining to the medical drama genre.

BIBLIOGRAPHY

Aly, Mohamed (2005). *Survey on Multiclass Classification Methods.* Technical Report. Pasadena: California Institute of Technology.

Bird, Steven, Ewan Klein and Edward Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* Sebastopol: O'Reilly Media, Inc.

Bradski, Gary (2000). "The OpenCV Library." *Dr. Dobb's Journal of Software Tools* 120: 122-125. https://www.drdobbs.com/open-source/the-opencv-library/184404319 (last accessed 16-07-23).

Buch, Shyamal, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei and Juan Carlos Niebles (2022). "Revisiting the 'Video' in Video-Language Understanding." In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2917-2927. Washington: IEEE Computer Society. https://doi.org/10.1109/CVPR52688.2022.00293.

Clark, Jeffrey A. (2015). "Pillow (PIL Fork) Documentation." *Version 10.0.0.* https://pillow.readthedocs.io/en/stable/ (last accessed 16-07-23).

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 8440-8451. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

Defazio, Aaron and Samy Jelassi (2022). "Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization." *Journal of Machine Learning Research* 23(1): 1-34.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171-4186. Minneapolis: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Du, Yifan, Zikang Liu, Junyi Li and Wayne Xin Zhao (2022). "A Survey of Vision-Language Pre-trained Models." In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence,* edited by Luc De Raedt, 5436-5443. Online: IJCAI. https://doi.org/10.24963/ijcai.2022/762.

Géron, Aurélien (2017). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol: O'Reilly Media, Inc.

Huang, Yu, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao and Longbo Huang (2021). "What Makes Multi-modal Learning Better than Single (Provably)." In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, edited by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan, 10944-10956. La Jolla: Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Kiela, Douwe, Suvrat Bhooshan, Hamed Firooz, Ethan Perez and Davide Testuggine (2019). "Supervised Multimodal Bitransformers for Classifying Images and Text." *arXiv preprint.* https://doi.org/10.48550/arXiv.1909.02950.

Krippendorff, Klaus (1980). *Content Analysis: An Introduction to its Methodology.* Newbury Park: Sage Publications.

Lei, Jie, Tamara Berg and Mohit Bansal (2022). "Revealing Single Frame Bias for Video-and-Language Learning." *arXiv preprint.* https://doi.org/10.48550/arXiv.2206.03428.

Li, Linjie, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng et al. (2021). "Value: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation." *arXiv preprint.* https://doi.org/10.48550/arXiv.2106.04632.

Liu, Chang, Armin Shmilovici and Mark Last (2023). "MND: A New Dataset and Benchmark of Movie Scenes Classified by Their Narrative Function." In *Proceedings of the 17th European Conference on Computer Vision*, edited by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, 610-626. Cham: Springer. https://doi.org/10.1007/978-3-031-25069-9_39.

Liu, Jingzhou, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang and Jingjing Liu (2020). "Violin: A Large-scale Dataset for Video-and-Language Inference." In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10900-10910. Washington: IEEE Computer Society. https://doi.org/10.1109/CVPR42600.2020.01091.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled Weight Decay Regularization." *arXiv preprint.* https://doi.org/10.48550/arXiv.1711.05101.

Muti, Arianna, Katerina Korre and Alberto Barrón-Cedeño (2022). "UniBO at SemEval-2022 Task 5: A Multimodal Bi-Transformer Approach to the Binary and Fine-grained Identification of Misogyny in Memes." In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, edited by Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, 663-672. Seattle: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.91.

Nagrani, Arsha, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid and Andrew Zisserman (2020). "Speech2action: Cross-Modal Supervision for Action Recognition." In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10317-10326. Washington: IEEE Computer Society. https://doi.org/10.1109/CVPR42600.2020.01033.

Neskorozhenyi, Rostyslav (2021). "How to get high score using MMBT and CLIP in Hateful Memes Competition." *Towards Data Science.* https://towardsdatascience.com/how-to-get-high-score-using-mmbt-and-clip-in-hateful-memes-competition-90bfa65cb117 (last accessed 16-07-23).

Pescatore, Guglielmo and Marta Rocchi (2019). "Narration in Medical Dramas I. Interpretative Hypotheses and Research Perspectives." *La Valle dell'Eden* 34: 107-115.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. (2021). "Learning Transferable Visual Models from Natural Language Supervision." *Proceedings of Machine Learning Research* 139: 8747-8763.

Rocchi, Marta and Guglielmo Pescatore (2022). "Modeling Narrative Features in TV Series: Coding and Clustering Analysis." *Humanities and Social Sciences Communications* 9(1): 1-11. https://doi.org/10.1057/s41599-022-01352-9.

Shen, Sheng, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao and Kurt Keutzer (2021). "How Much can CLIP Benefit Vision-and-Language Tasks*?" arXiv preprint.* https://doi.org/10.48550/arXiv.2107.06383.

Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy and Cordelia Schmid (2019). "VideoBERT: A Joint Model for Video and Language Representation Learning." In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 7464-7473. Washington: IEEE Computer Society. https://doi.org/10.1109/ICCV.2019.00756.

Tapaswi, Makarand (2016). *Story Understanding through Semantic Analysis and Automatic Alignment of Text and Video.* PhD dissertation. Karlsruhe: Karlsruhe Institute of Technology.

Tapaswi, Makarand, Martin Bäuml and Rainer Stiefelhagen (2015). "Aligning Plot Synopses to Videos for Story-based Retrieval." *International Journal of Multimedia Information Retrieval* 4(1): 3-16. https://doi.org/10.1007/s13735-014-0065-9.

Wei, Yixuan, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen and Baining Guo (2022). "Contrastive Learning Rivals Masked Image Modeling in Fine-Tuning via Feature Distillation." *arXiv preprint.* https://doi.org/10.48550/arXiv.2205.14141.

Weng, Zejia, Lingchen Meng, Rui Wang, Zuxuan Wu and Yu-Gang Jiang (2021). "A Multimodal Framework for Video Ads Understanding." In *Proceedings of the 29th ACM International Conference on Multimedia*, edited by Heng Tao Shen and Yueting Zhuang, 4843-4847. New York: Association for Computing Machinery. https://doi.org/10.1145/3474085.3479202.

Wu, Zuxuan, Caiming Xiong, Chih-Yao Ma, Richard Socher and Larry S. Davis (2019). "Adaframe: Adaptive Frame Selection for Fast Video Recognition." In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 1278-1287. Washington: IEEE Computer Society. https://doi.org/10.1109/CVPR.2019.00137.

Zahiri, Sayyed and Jinho Choi (2017). "Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks." *arXiv preprint.* https://doi.org/10.48550/arXiv.1708.04299.

Zhu, Linchao and Yi Yang (2020). "ActBERT: Learning Global-Local Video-Text Representations." In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8746-8755. Washington: IEEE Computer Society. https://doi.org/10.1109/CVPR42600.2020.00877.

TOWARD THE AUTOMATIC IDENTIFICATION OF ISOTOPIES

In Stefania Antonioni and Marta Rocchi (eds). *Investigating Medical Drama TV Series: Approaches and Perspectives. 14th Media Mutations International Conference*. Bologna: Media Mutations Publishing. https://doi.org/10.21428/93b7ef64.8ac7a6ca.