Identifying medical drama narratives: A multi-output regression approach with CNNs and BERT

Alice Fedotova
Dept. of Interpreting and Translation
University of Bologna, Forlì
alice.fedotova@studio.unibo.it
February 2, 2023

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 research has indicated that human coding is more effective in handling more complex variables at the core of communication studies (Hase, 2022), audiovisual content is often neglected because analyzing it is difficult and time-consuming. In this preliminary work, the task of analyzing narratives in medical dramas is addressed by leveraging spoken dialogues in the form of subtitles. The corpus was obtained by performing a temporal alignment between the subtitles and the dataset presented in Rocchi and Pescatore (2022), which consists of video segments labeled with weights from 0 to 6 representing the distribution of three plot types. Three text regression models were developed in order to approach this task. The best-performing model, based on BERT, obtained an R^2 score of 0.08 This can be considered a positive outcome, given the complexity of the regression task and the limitations of relying on a single modality.

1 Introduction

Medical dramas are a popular genre of television programming that depicts the lives and work of medical professionals, such as doctors, nurses, and other healthcare workers. These shows often follow the personal and professional lives of the main characters, as well as their interactions with patients and colleagues. Medical dramas have been a staple of television for decades, with some of the most well-known examples including “Grey’s Anatomy,” “ER,” and “House.” In light of the recent public health emergency, the relevance of this genre has become even more pronounced (Rocchi, 2019).

Currently, medical dramas are being analyzed within the framework of narrative ecosystems, a theoretical perspective that has been proposed for the investigation of “vast audiovisual narratives”, i.e. TV shows that are characterized by the need to maintain an “ongoing structure with narrative consistency and thematic coherence throughout large numbers of episodes and sometimes seasons” (Pescatore and Innocenti, 2017). According to Pescatore et al. (2014), modern TV series tend to feature multiple storylines and a high level of narrative complexity: this kind of audiovisual narratives requires a shift in thinking from the traditional concept of a “text," which is typically closed and limited in time and space, to the idea of a “narrative ecosystem," an open system which is similar to a natural environment, connecting narrative-textual, production and consumption dynamics (Rocchi, 2019).

In the field of media studies, content analysis has long been employed as a 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, 1995). Previous research has identified three fundamental categories or “isotopies" that characterize the medical drama genre: the medical case, which refers to the new medical cases that are presented in each episode; the professional plot, which focuses on the portrayal of medical professionals and their work; and the sentimental plot, which focuses on the personal lives and relationships of the characters (Rocchi and Pescatore, 2019).

Content analysis has been used to investigate medical dramas by assigning isotopies to segments, i.e. “portions of video characterized by space-time-action continuity” (Rocchi and Pescatore, 2022). This poses a challenge for automated approaches to content analysis, as modern segmentation algorithms are not efficient

at identifying homogeneous units that are relevant to the identified isotopies. Additionally, the process of coding requires a significant degree of expert knowledge and extensive training of the annotators performing the analysis. The task of manually identifying segments that pertain to the isotopies, combined with the time required to proficiently code, makes content analysis a significantly time-consuming process (Rocchi and Pescatore, 2022).

Consequently, this study aims to address the task of predicting the narrative composition of a TV drama with a set of multioutput regression models in order to analyze the dialogues at the subtitle level. The proposed architectures learn to predict both the plot types contained in a subtitle and the amount of time dedicated to each narrative, expressed with a continuous value from 0 to 6. The best-performing model, based on BERT, obtained an R^2 score of 0.08. While visual and audio information has been widely studied, textual clues are less explored for video understanding (Weng at al., 2021). The results point to the need to leverage other information in addition to the subtitles, which has proven beneficial for video understanding tasks (Li et al., 2021; Liu et al., 2020). This preliminary study constitutes a baseline for future work in the field of machine learning for audiovisual content analysis.

2 Dataset Construction

2.1 Background

The present work builds upon the dataset outlined in Rocchi and Pescatore (2022), created in the context of the project "NEAD framework. A systemic approach to contemporary serial product. The medical drama case."[1] The dataset includes more than 400 hours of video and consists of eight North American medical dramas, for a total of 32 seasons and 608 coded episodes. Episodes were coded by following a three-step content analysis protocol. First, three isotopies underlying the medical drama genre were identified: the medical cases plot (MC), the professional plot (PP), and the sentimental plot (SP). The medical cases plot 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 deals with the relationships and dynamics within the hospital among doctors and other medical staff. Lastly, the sentimental plot 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 (Rocchi and Pescatore, 2022).

The second step involved breaking down each episode into segments, which are defined as the units of the audiovisual product that possess continuity in terms of space, time, and action, as well as consistency in terms of thematic and narrative elements. For each segment, start and end times were identified and recorded for future analysis. This aspect is especially important, as it allowed the subsequent alignment with the text of the subtitles (see section 2.2). The actual coding phase followed the identification of the segments. During this phase, the appropriate isotopies were assigned to 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 received a weight of 0 (Rocchi and Pescatore, 2022).

2.2 Data Alignment

The availability of start times and end times 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: 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.

[1] https://dar.unibo.it/en/research/research-projects/prin-narrative-ecosystem-analysis-and-develompment-framework-nead-framework-un-approccio-sistemico-al-prodotto-seriale-contemporaneo-il-caso-del-medical-drama

Subtitle	PP	SP	MC
amelia invited riggs to dinner at our house.	0	1	0
it's about the whole healthcare system, not this place.	1	0	0
noelle webb, 43, complains of abdominal pain	0	0	1
i'm the chief of general, i loved working with you,	0.5	0.5	0
if I have to look him in the eye and tell him i blew it...	0	0.83	0.16
yes, well, the medical community and i are in a fight.	0.33	0.66	0
why? my patient is terrified.	0.33	0	0.66
yeah, not by you. page surgery.	0.66	0	0.33
whatever. he bends his rules all the time to save his own patients.	0.5	0.5	0

Table 1: Some instances from the dataset, after preprocessing (lowercasing and target normalization).

By relying on these features, the SRT files were processed in order to extract the timestamps and the text of the subtitles. The Pandas[2] library was then used to transform the start times and end times into DateTime objects, enabling operations on dates and timestamps.

For the purpose of aligning the subtitles with the provided dataset, a method for assigning each of the subtitles to the corresponding segment was identified. Inspired by Tapaswi et al. (2014) in which subtitles occurring at video shot boundaries were assigned to the shot which has a majority portion of the subtitle, the average 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 average 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. A dataset containing subtitles labeled with the corresponding narrative lines was therefore obtained by repeating the outlined procedure on the data relative to five seasons from the show Grey's Anatomy, for a total of 106 episodes and 101,685 subtitles. Table 1 illustrates some instances from the resulting dataset, with an example for each possible narrative or combination of narratives.

2.3 Data Partitioning

A hold-out test partition was used to evaluate the performance of the models. This method involves dividing the data into a training set and a separate test set, where the models are trained on the former set and then evaluated on the latter. A validation set was also used for the purpose of optimization. This approach was chosen over k-fold cross-validation as it was found to be too computationally expensive to perform in the conducted experiments. The hold-out method allows for a more efficient evaluation of the models while still providing a reliable measure of their performance. Table 2 shows the total number of instances per partition.

	PP	SP	MC	Total
Train	17,391	34,555	25,250	77,196
Dev	3,773	7,343	5,415	16,531
Test	3,701	7,384	5,449	16,534
Overall	24,865	49,282	36,124	110,271

Table 2: Counts intended as non-zero instances: SP is the most prevalent in the case of Grey's Anatomy.

3 Models and Representations

3.1 CNNs

Experiments were first conducted with two CNNs. As they are meant to be read and understood quickly, subtitles have a fixed size and are usually confined to a specific area of the screen. CNNs are designed to take advantage of the spatial structure within the data, which makes them particularly effective at processing sequences that have a consistent length. As for the text representations, the embeddings for the CNNs were obtained using GloVe after lowercasing and tokenizing the text with the Keras tokenizer. In the conducted experiments, GloVe was found to be less computationally expensive than word2vec while maintaining a comparable level of performance.

[2] https://pandas.pydata.org/

Parameters	Settings	CNN_1	CNN_2
Conv1D kernel size	∈ [1, 2, 3, 4, 5]	3	2
Conv1D filters	∈ [50 … 250] with a step size of 25	250	50
Dense units	∈ [32 … 256] with a step size of 32	250	352
Dropout value	∈ [0.05, 0.1, 0.2, 0.3]	0.2	0.1
Intermediate layers	∈ [1, 2, 3]	1	2
Loss function	MAE, MSE	MSE	MSE
Adam learning rate	∈ [0.002, 0.001, 0.0001]	0.001	0.001

Table 3: Manually selected hyperparameters (CNN_1), compared with the hyperparameters selected by HyperBand (CNN_2).

The first CNN was built using a CNN layer with 250 filters and a kernel size of 3, global max-pooling, a fully connected layer with 250 neurons, a dropout of 0.2, and a ReLU activation function. A sigmoid was used as the last layer's activation function, as the three plot types are not mutually exclusive (e.g., some instances can be assigned to more than one narrative, therefore having two or even three non-zero labels). As a consequence, target values were normalized between 0 and 1 before training. This ensured that predictions would never be less than 0 or greater than 1. The model was trained for eight epochs with a batch size of 32, although it stopped improving after the second. Consequently, the Model Checkpoint callback saved the model trained over two epochs.

The second CNN was created by optimizing the hyperparameters of the first one. For this purpose, the Keras Tuner [3] library was employed. The following hyperparameters were explored: kernel size, number of filters, dense layer units, number of dense layers with dropout, dropout value, learning rate, and loss function. The HyperBand algorithm[4] was used in order to explore potential configurations. The search was conducted over 20 epochs in conjunction with Early Stopping, a technique that consists in specifying an arbitrarily large number of training epochs and then interrupting the training phase when the model does not show any improvement. Furthermore, the patience parameter was set as 5, which is the number of epochs to wait before stopping if no progress is made. The structure of the optimized CNN is illustrated in Table 3. The identified hyperparameters were used to train a second CNN over five epochs with a batch size of 32. Any additional training resulted in overfitting.

3.2 BERT

Recent studies have demonstrated that pre-trained models like BERT can achieve state-of-the-art results in many natural language processing applications (Sun et al., 2019). Experiments were therefore conducted with the popular *bert-base-uncased model* in order to investigate the possibility of further improving the results achieved on this task by obtaining more sophisticated text representations. No additional preprocessing was performed apart from applying the BertTokenizer. A challenge that was encountered is that the HuggingFace Transformers library did not have a model available for a regression task with three outputs, unlike for classification tasks. This required the creation of a new model based on BERT, with the addition of a sigmoid layer with three outputs. The output layer was therefore identical to the one used in the CNN setting, so as to constrain the predictions within the [0, 1] range.

In order to fine-tune the transformer architecture, experiments were conducted according to the hyperparameters suggested by the authors of BERT (Devlin et al., 2019). Consequently, batch sizes of 16 and 32 were tested over 4 epochs with learning rates of 2e-5, 3e-5, and 5e-5 for the Adam optimizer. Overall, a learning rate of 2e-5 was more effective on this task than the other recommended values of 3e-5 and 5e-5. A batch size of 32 was also found to be preferable to that of 16, and the ideal number of epochs resulted to be two. Training for more epochs resulted in overfitting, which negatively impacted the performance of the model. As in the CNN setting, MSE was chosen as the loss function.

[3] https://keras.io/keras_tuner/
[4] https://keras.io/api/keras_tuner/tuners/hyperband/

	R^2 ↑				MAE ↓				RMSE ↓			
	PP	**SP**	**MC**	**All**	**PP**	**SP**	**MC**	**All**	**PP**	**SP**	**MC**	**All**
Baseline	0.00	0.00	0.00	0.00	0.26	0.39	0.37	0.34	0.32	0.43	0.41	0.15
CNN $_1$	-0.05	-0.05	-0.06	-0.05	0.24	0.38	0.35	0.32	0.33	0.44	0.42	0.16
CNN $_2$	0.02	0.05	0.03	0.03	0.26	0.37	0.36	0.33	0.32	0.41	0.41	0.15
BERT	**0.05**	**0.11**	**0.07**	**0.08**	**0.19**	**0.35**	**0.31**	**0.28**	**0.37**	**0.47**	**0.46**	**0.19**

Table 4: Performance of the experimented models on the test partition, evaluated with R^2, MAE, and RMSE.

4 Results and Evaluation

In order to evaluate the performance of the proposed architectures on this task, three metrics were examined: R-squared (R^2), mean absolute error (MAE) and root mean squared error (RMSE). According to Chicco et al. (2021), R^2 is more informative and truthful than other metrics like SMAPE and does not have the interpretability limitations of MSE, RMSE and MAE. It is a value between 0 and 1, with a higher value indicating a better fit of the model to the data. R^2 measures the performance of a model compared to a naïve model that always guesses $\bar{y}$, i.e. the mean. Values of R^2 outside the range 0 to 1 occur when the model fits the data worse than the mean. A baseline model which always predicts $\bar{y}$ has an R^2 of 0.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

MAE is the average absolute error between actual and predicted values, while RMSE is the square root of the mean squared error between the predicted and actual values. In both cases, the lower the value the more accurate are the predictions. However, RMSE penalizes large errors more than MAE due to the fact that errors are squared. This makes RMSE more sensitive to outliers compared to MAE. When it comes to regression problems, it is common to create baseline models that predict the mean or median of the training data output. The most trivial model minimizing squared error is predicting the mean for all the samples, whereas for mean absolute error, it is predicting the median.

Two of the proposed models outperformed the baseline by a small margin, even though the first CNN obtained an R^2 lower than 0. The BERT model has the best R^2 and MAE, but a slightly higher RMSE as well. This could be due to the fact that the model is able to find a solution that is different from simply predicting the mean, as evidenced by the improved R^2. As a consequence, it is possible that large errors occur more often. Looking solely at the MAE and RMSE, it is not clear what type of narrative the models perform best on. Compared to SP, the BERT model obtained a lower MAE and RMSE on PP. However, by looking at the R^2, the performance of the BERT model on SP is actually twice as good as on PP. This could be a consequence of the fact that the Grey's Anatomy data contained a higher-than-average SP component, meaning that the model had more information about this type of plot in the training phase.

5 Related Work

Videos, through their ability to simultaneously engage various human faculties such as hearing, sight, and emotions, are more effective in conveying narratives compared to static images or text. However, the addition of the time element through shots and scenes makes it extremely complex for machine learning models to understand the narrative content of a video. 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 the stories conveyed in videos, making them more searchable and accessible (Tapaswi, 2016). Despite the fact that there is yet no agreement on which modality is best when eliciting high-level meaning from audiovisual content, researchers believe that two or more modalities are better than one (Bayoudh, 2022).

Compared to visual and audio information, textual clues are less explored for video understanding (Weng at al., 2021) In the broader context of movies and TV shows, the 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). Most previous works only leverage visual information from the video channel (Miech et al., 2019; Zhu and Yang, 2020). However, as demonstrated by Li et al. (2021), leveraging both video and subtitle channels achieves the best performance on the VALUE benchmark, which includes 11 video understanding tasks from a variety of datasets and video genres. A similar result is reported by Liu et al. (2020) 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.

In the field of sentiment analysis, related work has been conducted on the TV show Friends. Zahiri and Choi (2017) similarly employ a CNN 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.

Subdividing emotions into a small set of emotion categories is typical of categorical approaches to emotion detection. However, dimensional approaches based on regression have emerged as well. Compared with emotion classification, emotion regression had a late start due to the lack of large-scale annotated corpora and the inherent difficulty of the regression task (Zhu et al., 2019). A popular model is the Valence-Arousal-Dominance (VAD) model, which is used in the EMOBANK corpus (Hahn and Buechel, 2016).

6 Conclusions and Future Work

As discussed, modeling the narrative structure of a TV drama is a non-trivial task. After all, "a narrative cannot be reduced to the sum of its sentences" (Lacey, 2000). The problem of predicting the narrative composition of Grey's Anatomy was addressed with three multioutput regression models in order to learn both the narratives and the time dedicated to each one of them, at the level of a subtitle. The best performing model, based on BERT, obtained an R^2 score of 0.08. Given the difficulty of this task, this is a promising result. Relying exclusively on the subtitles constitutes the main limitation of this study, which will be addressed in the future by including information from the video channel as well. It should be also noted that modeling the relationship between the subtitles and the narratives constitutes a significantly fine-grained task, which was motivated by the fact that subdividing the text into higher-order units such as segments would require human intervention. Ideally, a model should be able to generalize to other unseen episodes. This means that the only information available to the model would be individual subtitles, as extracted from a subtitle file. In the future, this issue will be addressed once an acceptable level of accuracy is reached. If a model manages to predict the narratives reasonably well, further experiments will be conducted in order to assess the possibility of performing segmentation based on a sliding window that detects changes in the narrative composition according to the predictions of the model.

References

Bayoudh, K., Knani, R., Hamdaoui, F., &; Mtibaa, A. (2021). A survey on Deep Multimodal Learning for Computer Vision: Advances, trends, applications, and datasets. The Visual Computer, 38(8), 2939–2970. https://doi.org/10.1007/s00371-021-02166-7

Buechel, S. & Hahn, U. (2016). Emotion analysis as a regression problem: dimensional models and their implications on emotion representation and metrical evaluation. Proceedings of the Twenty-second European Conference on Artificial Intelligence. https://doi.org/10.3233/978-1-61499-672-9-1114

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7. https://doi.org/10.7717/peerj-cs.623

Hase, V. (2022). Automated Content Analysis. In F. Oehmer-Pedrazzi, S. H. Kessler, E. Humprecht, K. Sommer & L. Castro (Eds.), Standardized Content Analysis in Communication Research (pp. 23-36). Springer. https://doi.org/10.1007/978-3-658-36179-2_3

Innocenti, V and Pescatore, G. (2017). Narrative Ecosystems. A Multidisciplinary Approach to Media Worlds. In M. Boni (Ed.), World Building. Transmedia, Fans, Industries (pp. 164-184).

Amsterdam University Press. https://doi.org/10.1515/9789048525317-010

Krippendorff, K. (2019). Content analysis: An introduction to its methodology. SAGE.

Lacey, N. (2000). Key concepts in media studies. Macmillan.

Li, L., Lei, J., Gan, Z., Yu, L., Chen, Y. C., Pillai, R., ... & Liu, Z. (2021). Value: A multi-task benchmark for video-and-language understanding evaluation. arXiv preprint arXiv:2106.04632

Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., & Liu, J. (2020). Violin: A large-scale dataset for video-and-language inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10900-10910).

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2630-2640).

Nagrani, A., Sun, C., Ross, D., Sukthankar, R., Schmid, C., & Zisserman, A. (2020). Speech2action: Cross-modal supervision for action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10317-10326).

Pescatore, G. and Rocchi, M. (2019) Narration in Medical Dramas: Interpretative Hypotheses and Research Perspectives. La Valle dell'Eden, 1, 107-115.

Pescatore, G., Innocenti, V., & Brembilla, P. (2014). Selection and evolution in narrative ecosystems. A theoretical framework for narrative prediction. 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). https://doi.org/10.1109/icmew.2014.6890658

Rocchi, M. (2019). History, Analysis and Anthropology of Medical Dramas: A Literature Review. Cinergie - Il Cinema e le Altre Arti, 8(15), 69-84. https://doi.org/10.6092/issn.2280-9481/8982

Rocchi, M. and Pescatore, G. (2022). Modeling Narrative Features in TV series: Coding and Clustering Analysis. Humanities and Social Sciences Communications, 9(333), 1-11. https://doi.org/10.1057/s41599-022-01352

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18 (pp. 194-206). Springer International Publishing.

Tapaswi, M. (2016). Story understanding through semantic analysis and automatic alignment of text and video [Doctoral dissertation, Karlsruhe Institute of Technology]

Tapaswi, M., Bäuml, M., & Stiefelhagen, R. (2014). 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

Weng, Z., Meng, L., Wang, R., Wu, Z., & Jiang, Y. G. (2021). A Multimodal Framework for Video Ads Understanding. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 4843-4847).

Zahiri, S. M., & Choi, J. D. (2018). Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In Workshops at the thirty-second AAAI conference on artificial intelligence.

Zhu, L., & Yang, Y. (2020). Actbert: Learning global-local video-text representations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8746-8755).

Zhu, S., Li, S., & Zhou, G. (2019). Adversarial attention modeling for multi-dimensional emotion regression. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1045