7. Adopting NLP Techniques to Analyze Twitter Social Discourses around *The Good Doctor*

Greta Iapalucci

ABSTRACT

The Good Doctor (ABC, 2017-) is a US medical drama centered upon the life of Shaun Murphy, a surgeon with Autism Spectrum Disorder (ASD) and savant syndrome. While the serial product has been widely investigated from a linguistic and a medical humanities approach, little has been said about its reception. Based on this lack of research and considering the ability of medical dramas to promote information about health issues, the aim of this chapter is twofold: first, to analyze Twitter social discourses around *The Good Doctor* with a focus on the representation of ASD and, second, to explore the potential of Natural Language Processing (NLP) techniques to investigate the reception of TV series on Twitter. Two different methodologies were implemented to analyze the tweets: a topic modeling technique to detect the main topics, carried out using BERTopic, and a sentiment analysis approach to mine the audience's opinion on a subset of tweets related to ASD, conducted using RoBERTa. Findings have shown that that *The Good Doctor*'s audience mostly discuss narrative and fruition-related topics, specifically throughout its airing time period, and have mainly positively received the representation of ASD contained in it.

KEYWORDS

Medical drama; social discourses; ASD; topic modeling; sentiment analysis.

Introduction

The Good Doctor (ABC, 2017-) is a US medical drama created by David Shore and based on *Good Doctor* (KBS2, 2013), a South Korean series. It is centered upon the sentimental and professional life of the main character Shaun Murphy, played by Freddie Highmore, a young surgeon with Autism Spectrum Disorder (ASD) and savant syndrome.

As will be more thoroughly explained the second paragraph, the audio-visual serial product has been mainly investigated so far from linguistic (e.g., Aulia et el. 2022, Putra et al. 2019) and medical humanities (e.g., Moore 2019, Hilsabeck 2022) lenses, especially regarding the conditions of the protagonist. However, to the best of our knowledge, only one study has focused on its reception (Stern and Barnes 2019), yet limiting it to the ASD representation within the drama.

Based on this lack of research and taking as a starting point the ability of medical dramas to promote information about health issues and to raise awareness on sensitive topics (Abseek Brusse et al. 2015, Hether 2008), the aim of this chapter is twofold: first, to analyze Twitter social discourses around *The Good Doctor* with a close attention to the representation of ASD within it and, second, to explore the potential of quantitative data analysis tools, specifically Natural Language Processing (NLP) techniques, for the study of audiovisual products' reception.

More in detail, the first aim of this chapter can be broken down into two research questions (RQs), namely:

- RQ1: Which are the main topics around *The Good Doctor* discussed by the audience and how do they evolve over time?
- RQ2: What is the sentiment of the audience towards the representation of ASD within *The Good Doctor*?

The analysis was undertaken on Twitter, in line with much previous research that has focused on the reception of TV series (e.g., Andy et al. 2022,

Molteni and De Léon 2016). The remainder of the chapter is organized as follows: first, we will review previous works that have investigated *The Good Doctor*; second, we will focus on the data collection and on the methodology used to undertake the research; third, we will present the main results of the research, and, lastly, we will discuss them and outline future possible research lines.

The Good Doctor: Research Perspectives

So far, *The Good Doctor* has been primarily investigated from two research perspectives, namely linguistic and medical humanities.

On the one hand, beginning by reviewing the research articles that have taken a linguistic perspective, Koh (2018) adopted a corpus linguistic approach to put forward a comparison between the Korean original drama *Good Doctor* and the US adaptation, discovering that the latter one shows a higher lexical density than the former. This suggests that US screenwriters employ a wider vocabulary, with fewer repetitions, compared to the Korean ones.

Elaborating on the kind of vocabulary adopted in the TV series, Aulia et al. (2022) conducted a study to investigate medical terminology through semi-structured interviews and an open-ended questionnaire to 31 participants: an English teacher and 30 Indonesian-speaking medical students. The research highlighted not only that students feel the necessity to learn medical vocabulary-specific terms in English, but also that audiovisual objects which have a wide lexical variety like *The Good Doctor* could be useful teaching media in English for Specific Purposes (ESP) classes. The lexical variety of the medical drama under investigation has been further stressed by Tyasrinestu and Ardi (2020), who have focused on idiomatic expressions with the double goal of identifying the ones contained in *The Good Doctor* through a content analysis of the subtitles and studying the translation strategies of the above-mentioned expressions into the Indonesian translation of the drama. Findings revealed that many types of idiomatic expressions are present in *The Good Doctor* and acknowledged equivalence as the most frequent translation strategy.

Further linguistic research on *The Good Doctor* has adopted a pragmatic approach: Putra et al. (2019), for instance, have classified subtitles into

speech acts, finding that expressive speech acts are the most represented category due to the high presence of feelings and emotion within the dialogues, while Mahdi Mosin (2021), instead, focused on evidentiality in doctors' talk to express credibility and reliability, observing that they mainly use objective terms to carry out patients' diagnoses.

Enriching the study of the TV series from a linguistic perspective, three works have dealt specifically with the language disorder representation of a person with ASD through the character of Shaun Murphy, adopting either a pragmatic (Larasati 2019, Rokhim 2022) or a psycholinguistic approach (Dwiyanti 2022). In detail, Rokhim (2022)'s research contemplated 12 utterances elicited by Shaun and found out mainly examples of stuttering, followed in frequency by phonological disorder, difficulty in understanding the concept and in following the direction. Moreover, all kinds of language impairments found in his speech were experienced in strong psychological contexts for the character. Larasati (2019) broadens the scope of the aforementioned analysis by considering more dialogues and by investigating Shaun's both verbal and non-verbal language, observing many cases of verbal language disorder in his discourse (e.g., violation of the maxims of conversation, turn-taking violation, echolalia etc.) but also of non-verbal language disorder (e.g., limited use of gesture, inappropriate facial expression, stiff gaze etc.). The researcher also justified the linguistic impairments based on the social impairments of the character, namely lack of comprehension, difficulty at adjusting the tone of voice and lack of mind-reading skills. Dwiyanti (2022) focused on the language disorder of the character as well, but referring to the types of existing language disorders, i.e. expressive and receptive, and the kinds of ASD responses given. Through a descriptive qualitative research, findings showed that Shaun presented both types of language disorders, but the dominant one was receptive, while the most frequent ASD response given by Shaun was classified as cognitively-relevant.

On the other hand, many studies have focused on *The Good Doctor* from a medical humanities viewpoint. To begin with, two papers (Cambra-Badii and Baños 2018, Cambra-Badii et al. 2021) published within the field of bioethics should be acknowledged. Both contributions begin the analysis from the description of the representation of ASD and savant syndrome within the medical drama, in order to understand whether its presence could be useful to teach bioethics and consequently the doctor-patient communication to medical students. While the first contribution (Cambra-Badii

and Baños 2018) focuses more on the description of the representation of ASD within the TV series, the second (Cambra-Badii et al. 2021) draws from these premises to detect the situations in which a bioethical principle was involved through a content analysis. Because the analysis has identified many of these situations in the medical drama, the authors argue for its usefulness in bioethical and health education.

The representation of ASD within the series has been further investigated within academia. While Stark (2020)'s approach draws both from Foucauldian discourse analysis and an autoethnographic reflection and concludes that the medical drama is ambiguous because of its decision to represent only the controlled features of ASD, the majority of the studies were more critical. For instance, Kluke (2020) explores the high social competence of the character and argues about its inaccuracy, which can lead to create false expectations towards people with ASD among the audience.

Furthermore, Moore (2019) and Hilsabeck (2022) focus on the character's savant syndrome. The first work harshly criticizes the function of the savant syndrome within the TV series, because it is exploited as a means to enrich and reinforce the lives of neurotypical characters by acclaiming their behaviors and relationships and thus making people with ASD feel "not fully human" (Moore 2019: 300). The second work, instead, stresses that what makes Shaun a "good doctor" is precisely his savant syndrome, because it adds to the human side of his behavior a "machine" component, which contributes in emphasizing his qualities but also in outdistancing him from his colleagues (Hilsabeck 2022: 80).

If the representation of ASD and, accordingly, of the savant syndrome in *The Good Doctor* has been mainly criticized within academia, little has been said about the reception of the audiovisual product, either in general nor in particular about the representation of the two related conditions. To the best of our knowledge, the only exception is Stern and Barnes (2019)'s work. In their study, the authors compared the effects of the exposure to one episode of the drama to those of one lecture about ASD through a survey submitted to 144 undergraduate students. They found out that the reception of the audiovisual product was widely more positive than the lecture: it led to a more accurate knowledge of ASD and simultaneously to a higher desire to know more about it.

Data and Methodology

As many previous studies that have investigated social discourses and TV series (e.g., Andy et al. 2022, Molteni and De Léon 2016), Twitter has been chosen as the source of all the datasets for this research on the account of its primarily textual component compared to other social networks. The data collection and pre-processing stages were all carried out using R. In detail, tweets were collected through *academictwitteR*, a R package to access Twitter API. A hashtag-based query was undertaken using *#thegooddoctor*, the hashtag both promoted by the official account of the TV series on social networks and the most used by Twitter users. The query was limited to the airing time period that spans from the day of the first aired episode of the first season (that is, September 25, 2017) to a week after the conclusion of the last ended season at the time of undertaking this research, i.e. season 5 (that is, May 23, 2022). The collected dataset is composed of 385,826 tweets, which will be referred to hereafter as *Original TGD[1] dataset*. In order to have a homogeneous corpus and to reduce noise, only English tweets were used in this research. Accordingly, a new dataset was created, named *English TGD dataset* and composed of 276,079 tweets.

Two different methods were implemented to analyze social discourses: BERTopic for RQ1 and RoBERTa, a Robustly Optimized BERT Pretraining Approach, for RQ2. Both tools are based on BERT, Bidirectional Encoder Representations from Transformers (Devlin et al. 2019), a language model based on Transformers[2] that has become in few years a baseline in NLP. The whole analysis was conducted by means of Python.

Topic Modeling: BERTopic

In order to address RQ1, a topic modeling approach was performed, which automatically analyzes large collections of textual information by detecting the latent topics contained in them. BERTopic (Grootendorst 2022) was preferred over more popular topic models such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) based on

---

[1]   *The Good Doctor.*

[2]   Transformers are neural network architectures proposed by Vaswani et al. (2017) that have highly outperformed other models in machine learning and in NLP since their introduction. To know more about the adoption of Transformer in NLP, please see: Tunstall et al. (2022).

previous works that have proven that this transformer-based model that takes BERT as an embedder achieves greater state-of-the-art results on microblogging texts like tweets (Egger and Yu 2022).

To create the BERTopic model, a minor textual pre-processing stage of the *English TGD dataset* was needed. In detail, URLs (identified by the *http* beginning), mentions (beginning with the @ symbol), hashtags (beginning with the # symbol) were removed and only alphabetical letters were maintained, aimed to remove not only numbers and punctuation but also emojis. No further preprocessing action was made given that BERTopic has been shown to perform best on slightly preprocessed datasets (Grootendorst 2022: 6).

Short tweets were dropped out in order to minimize the presence of empty or noisy tweets. Hence, tweets that contained less than 30 characters were removed. A new dataset composed only of long tweets was created, the *Long English TGD dataset*, amounting to 198,997 tweets.

In order to fit the BERTopic model on the *Long English TGD dataset*, 5 hyperparameters need to be tuned, corresponding to the 5 preliminary steps to create the topic model, namely:

1. **Generating embeddings**: this step consists in the transformation of the input documents[3] into numerical representations. The default BERTopic pre-trained embedding model was kept to perform this task, which is *sentence-transformer-MiniLM-L6-v2*, an English language model trained for semantic similarity tasks.

2. **Reducing dimensionality**: after having created the numerical representations of the documents, it is crucial to reduce their dimensions, given that the cluster models of the third step perform better on low-dimensional data. While the default dimensionality reduction algorithm in BERTopic is UMAP (Uniform Manifold Approximation & Projection), for the present study PCA (Principle Component Analysis) was adopted, on the account of its ability to better reduce the dimensions.[4] Additionally, the output dimension, that is the dimension of the data to be passed to the clustering mode, was set to 5, to maximize the reduction of the dimensions.

3. **Clustering the reduced embeddings**: BERTopic default algorithm for embedding clustering was used, i.e. HDBSCAN (Hierarchical

---

[3]   In this case, the *Long English TGD dataset* tweets.

[4]   As a reference, please visit: https://grabngoinfo.com/hyperparameter-tuning-for-bertopic-model-in-python/ (last accessed 01-07-23).

Density-Based Spatial Clustering of Applications with Noise). The reason for keeping it is twofold: first, it automatically identifies the clusters, without setting an a priori number, and, second, it creates a separate cluster for outliers (labeled as -1), thus reducing noise within the main clusters. The hyperparameter of a minimum cluster size was given to the model, which is the minimum number of documents that a cluster can contain, and it was set to 50 to reduce the resulting number of clusters yet at the same time preserving interpretability. As a consequence of the embedding clustering, documents are clustered into semantically similar clusters.

4. **Tokenizing the topics and removing stopwords**: the default vectorizer model, namely CountVectorizer, was kept to both tokenize the clusters and to load the English stopwords from.

5. **Extracting topic representation**: the default model, i.e. the c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) was maintained: this model's task is to convert the formed clusters into a single document and to extract the term frequency of words from the specific cluster, altogether computing the importance of each term within a class. This process defines the most representative words per each topic.

The BERTopic model tuned with these hyperparameters yielded 42 topics and 179,876 tweets as outliers (90.4% of the whole *Long English TGD dataset*).

To reduce the number of outliers, BERTopic default function was selected, which calculates the c-TF-IDF representations of outlier documents and re-assigns them to the best fitting representation of the topics. As a result, the redistribution of the outlier tweets led to a new BERTopic model with 211 tweets classified as outliers (0.01%) and the rest of the *Long English TGD dataset* assigned to topics (198,786 tweets, 99.9%).

To reduce the number of topics, instead, three steps were carried out. First, an automatic reduction was called, which is performed by automatically merging the topics that were close to each other based on HDBSCAN results. This function generated a model of 26 topics. Second, two visualization functions, i.e. topic hierarchy and topic similarity, were called on the new updated model, in order to figure out how topics were related to one another and, eventually, whether a further reduction on the topics could be undertaken. The results of the two functions are visualized respectively in Figures 1 and 2. The topic hierarchy function is able to visualize the group-
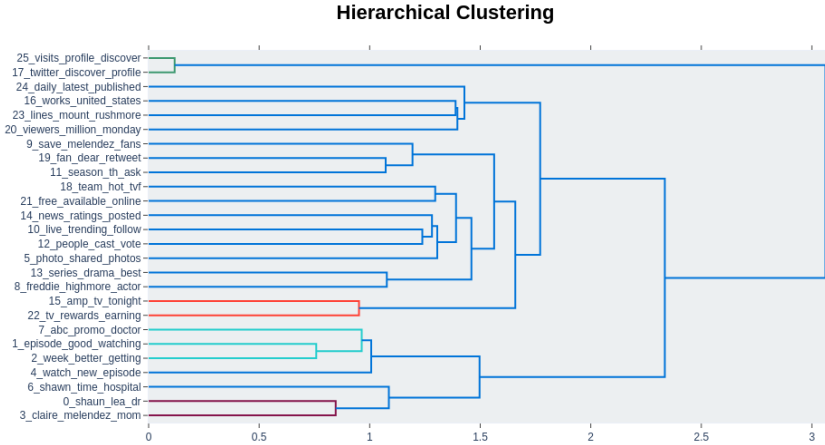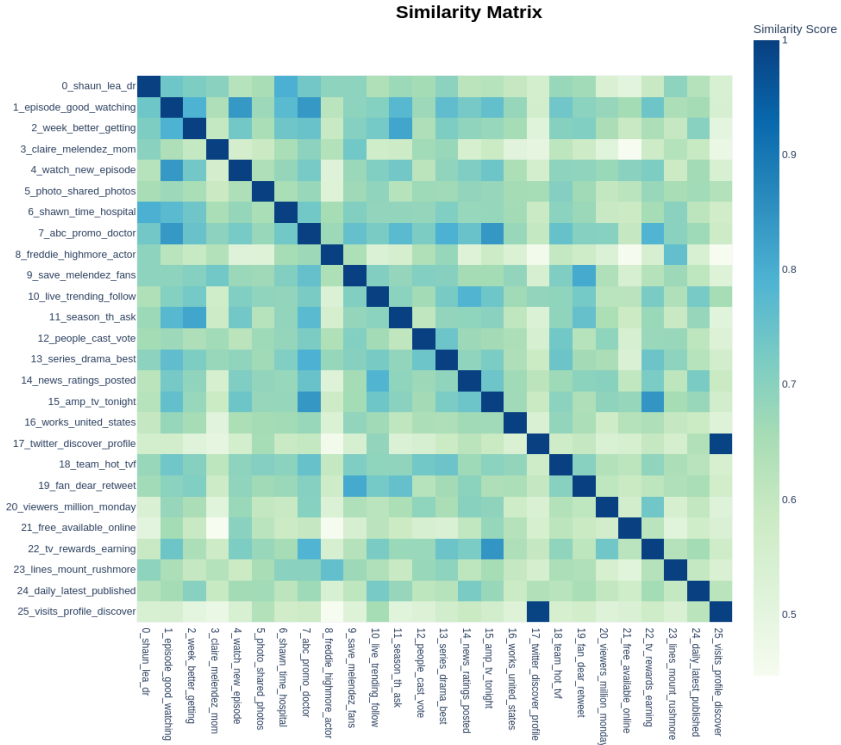
FIGURE 1
Dendrogram showing the hierarchical clustering of the topics.

ing of topics into a hierarchical structure displayed through a dendrogram and based on the distance matrix of the c-TF-IDF representation for each topic (Figure 1). Based on this hierarchy, a calculation of the topic representation at each merged step is carried out and the closer the merge between two topics is to 0, the less distant and thus more similar the topics are. Conversely, the farther their node is from 0, the more different the topics are. Topic similarity, instead, displays a heatmap of the similarity between topics (Figure 2). It first calculates cosine similarities between the created topic embeddings, whose range goes from 0 to 1, i.e., from no similarity to perfect similarity between topics, and then visualizes it through a heatmap, in which the darker the blue square resulting from the overlapping of two topics is, the more similar the topics.

Third, a manual merge of the topics was undertaken based on the results of these two visualization functions, taking as thresholds topics whose clustering node in the hierarchy was below 1 and that altogether had a similarity score higher than 0.70. This merge yielded to a further reduction of the model to 21 topics.

Heatmap showing similarity between topics.

Sentiment Analysis: RoBERTa

RQ2 was addressed through a sentiment analysis approach, which is the NLP task that classifies texts by extracting subjective information, that is the sentiment, and tagging it as either positive, negative or neutral. RoBERTa (Liu et al. 2019) was selected for this research, which is a language model based on BERT, yet optimized pertaining to the hyperparameters and training steps. The pre-trained twitter-RoBERTa-base model for sentiment analysis was adopted,[5] trained on more than 124 million tweets and fine-tuned according to the TweetEval benchmark (Barbieri et al. 2020).

---

[5]   The code of the model is available at https://huggingface.co/cardiffnlp/twitter-roberta -base-sentiment-latest (last accessed 01-07-23).

Aimed at mining the audience's sentiment towards the depiction of ASD within *The Good Doctor*, a subset from the *English TGD dataset* was created containing only tweets related to this specific topic, filtered through the use of hashtags by Twitter users. The *English TGD dataset* was first pre-processed by lowercasing the words, by standardizing the characters and accents and by leaving out punctuation markers: these steps were carried out in order to guarantee that similar hashtags but with differences in lowercase or uppercase letters (e.g., "#thegooddoctor" vs "#TheGoodDoctor"), in characters (e.g., "#thegooddoctor" vs "#thégòòddoctor") or in punctuation (e.g., "#thegood-doctor" vs "#thegoodoctor.") could not account for multiple hashtags.

Once having pre-processed the tweets, a list of all the hashtags comprised in the corpus was generated. In order to create a subset only of tweets that discussed ASD within *The Good Doctor*, the list of hashtags was closely scrolled down to identify hashtags related to ASD that occurred at least 10 times in the *English TGD dataset*. The resulting ASD-related hashtags are reported in Table 1.

ASD-related hashtags
#autism
#autismawareness
#autistic
#asd
#autismacceptance
#actuallyautistic
#autismawarenessmonth
#pmfautism
#aspergers
#autismmom
#worldautismawarenessday
#autismacceptancemonth
#autismspeaks
#savantsyndrome
#autismeducation
#autismawarenessday
#autismspectrum
#autismrocks
#autismo[6]

TABLE 1
List of ASD-related hashtags that occurred more than 10 times in the *English TGD dataset*.

---

[6] "Although the list of hashtags was extracted from the English TGD dataset, autismo is a not an English word. Therefore, tweets with this particular hashtag were manually

A new dataset was created with all the tweets that contained at least one of the hashtags present in Table 1 and was named the *ASD dataset*. Its size is 1,360 tweets. Before applying the RoBERTa sentiment analysis model to the *ASD dataset*, it is necessary to tokenize the tweets. To do so, the Hugging Face Transformers library provides AutoTokenizer, a class that can load a tokenizer from a pretrained model, twitter-RoBERTa-base-sentiment-latest in this case.

Once the model, the *ASD dataset*, the labels (i.e., positive, neutral and negative) and the tokenizer were provided, a function was created to apply the model to the *ASD dataset* in which the input is the tokenized text and the output required is one of the three pre-defined labels attached to every tweet, with a specific score from 0 to 1 for every label, identifying the probability distribution of the sentiment of every tweet over the three labels. Accordingly, the higher the score attributed to a tweet, the more confident the model is in the attribution of that label.

An overview of all the datasets used in this research and their respective size is reported in Table 2.

Dataset	Size
Original TGD dataset	385,826
English TGD dataset	276,079
Long English TGD dataset	198,997
ASD dataset	1,360

TABLE 2
Size of datasets.

checked to assess whether they were written in English. It was shown that the tweets were either in both English and Spanish or they did not contain any text but only several hashtags."

Topic Modeling Results

Frequency and Category of Topics

As a result of the methodology outlined in 3.1 and applied to the *Long English TGD Dataset*, the BERTopic model generated 21 topics. In order to analyze these topics, the most important words and the most representative tweet per each topic were displayed. Table 3 shows an overview of the yielded topics.

TABLE 3
List of topics with their topic ID and their frequency, the top 4 describing words and the most representative tweet per topic, in a decreasing order.

Topic ID	Count	Most describing words	Most representative tweet
-1	211	Fitzgerald, Larry, cowboys, Jerry	/
0	86,307	Shaun, Claire, Lea, dr	Lea took all that time to make pancakes for Shaun and he just goes and leaves like that?? #TheGoodDoctor
1	60,580	episode, good, doctor, season	I can't with this!! First episode and already crying! #thegooddoctor
2	9,372	watch, new, episode, tonight	I'm about to watch the new episode of #TheGoodDoctor.
3	6,624	photo, shared, photos, app	Shared from Photos app 1 photo @seale15eastcoast Giving it a try #FreddieHighmore #TheGoodDoctor #TGDsquad https://t.co/dlxpLP5Icy
4	6,074	Shawn, time, hospital, wonder	I have just watched episode 8 of Season 5 of The Good Doctor. I wonder what will be happening with Shawn Murphy, at the hospital next time. #TheGoodDoctor
5	4,469	Freddie, Highmore, actor, amazing	Wow, Freddie Highmore is an unbelievably amazing actor. #TheGoodDoctor @GoodDoctorABC
6	4,135	save, Melendez, fans, sign	The Good doctor fans: Save Melendez! #bringbackmelendez #TheGoodDoctor - Sign the Petition! https://t.co/MfWG7aOE45 via @UKChange

135

Topic ID	Count	Most describing words	Most representative tweet
7	3,259	live, trending, follow, rt	Currently trending scripted: 6. #SaveProdigalSon ðŸ"ª 7. #BetterCallSaul 8. #WynonnaEarp ðŸ © 9. #TWD 10. #TheGoodDoctor like, RT & follow https://t.co/23JilGU6Yb
8	3,178	season, th, ask, finale	We are fans of #TheGoodDoctor and we ask for the renewal of the show for one more season. Please @ABC_Publicity and @ABCNetwork renew @GoodDoctorABC for a 5th season #TheGoodDoctorSeason5 https://t.co/miVzSe5EYS
9	2,478	people, cast, vote, choice	I voted for #FreddieHighmore for #TheMaleTvStar of 2018 Cast your vote for the E! People's Choice Awards at #pcas! #TheGoodDoctor
10	2,099	TV, amp, tonight, listings	PlsRT @hillharper #TheGoodDoctor tonight on ABC. Free TV Listings @ https://t.co/BM9zp9GCSF &
11	1,853	series, drama, best, lead	For your #Emmys2021 consideration Outstanding Best Lead Actor in Drama Series #freddiehighmore #thegooddoctor #emmynominations #emmys @TheEmmys @GoodDoctorABC
12	1,726	news, ratings, posted, followed	Watch #TheGoodDoctor tonight at 10 on #WFTV followed by Eyewitness News at 11. https://t.co/OmtBzUasJi
13	1,401	Twitter, discover, profile, visits	Discover who visits your Twitter profile in the last 24h #ILikeBeer #TheBachelor #LHHMIA #DescribeYourselfBadly #TheGoodDoctor https://t.co/E2XHsWfyTo
14	1,153	works, United, States, seconds	Only works for the United States #TheBachelor #LHHMIA #DescribeYourselfBadly #ToddFrazier #TheGoodDoctor #LadyDoritos #PumpRules https://t.co/FwYSU6vfCF

Topic ID	Count	Most describing words	Most representative tweet
15	1,011	team, hot, TVF, online	HOT FROM TEAM TVF: Watch #TheGoodDoctor Online: Season 2 Episode 18 https://t.co/PTGo5yyvIT via @pauldailly1992
16	742	fan, dear, retweet, fans	CHAIN FAN PETITION!! Dear fans: copy the petition below and retweet "We are fans of #TheGoodDoctor and we ask for the renewal of the show for one more season. Please @ABC_Publicity and @ABCNetwork renew @GoodDoctorABC for a 5th season #TheGoodDoctorSeason5 â™¥ï¸" https://t.co/BVzFeTgYwl
17	675	viewers, million, Monday, ratings	Monday Ratings: #TheGoodDoctor = 5.6 million viewers; #AllAmerican = 772,000 viewers;#BlackLightning = 627,000 viewers; #911LoneStar = 5.5 million viewers and #Manifest = 3.8 million viewers.
18	662	free, available, online, watch	The Good Doctor is available to watch online for FREE! #TheGoodDoctor Watch now: https://t.co/1QKqKWBQfH https://t.co/vZNNz5L5p9
19	603	lines, Mount, Rushmore, freddiehighmore	#TheGoodDoctor 1x02 - Mount Rushmore - 5 Funniest Lines! https://t.co/CqnNtEZleN via @YouTube
20	385	daily, latest, published. dahlinke	The Dahlinke2306 Daily has been published https://t.co/28EGpRj9tW Vielen Dank an @davidhogg111 @maewald @DanRather #yr #thegooddoctor
TOTAL	198,997	/	/

Based on Table 3, a classification of the topics into 4 categories was carried out and is reported in Table 4.

As can be seen from Table 4, the narrative category is formed by only one topic, namely topic ID = 0, yet it constitutes the most frequent category among the four, representing the 43.4% of the tweets classified by the topic model. In detail, within the 4 most representing words, one can notice the names of three characters: Shaun, the protagonist, Claire, part of the main cast for the first four seasons, and Lea, the protagonist's first love interest

	Category of topics	Topic IDs	Count
1	Narrative	0	86,307 (43.4%)
2	Fruition	1, 2, 4, 7, 10, 12, 15, 17, 18, 19	86,061 (43.2%)
3	Fandom	5, 6, 8, 9, 11, 16	16,855 (8.5%)
4	Noise	3, 13, 14, 20	9,563 (4.8%)
	Outliers	-1	211 (0.1%)
	TOTAL	/	198,997 (100.0%)

Table 4

Classification of topics into 4 categories, ordered by the count of tweets belonging to each category.

and then fiancée. The most representative tweet of the topic focuses on the sentimental storyline between Shaun and Lea. The other most representative tweets, not reported here for space reasons, confirm the focus of this topic on the narration by recalling Shaun's ASD, or empathizing with him in an argument with dr. Lim, or Claire's patience. Other characters and respective storylines are cited, such as dr. Reznick and dr. Melendez.

The second most frequent category discusses the fruition of the TV series: it is composed of 10 topics, namely IDs = 1, 2, 4, 7, 10, 12, 15, 17, 18 and 19, and amounts to the 43.2% of the total, almost equaling the first category in frequency. Within this category, three sub-categories of topics can be identified, namely: topics related to the fruition of the TV series in general (IDs = 1, 2 and 4), topics connected to its synchronous fruition (IDs = 7, 10, 12 and 17) and topics related to its asynchronous fruition (IDs = 15, 18 and 19). Considering first the three topics related to the general fruition of the TV series, while in topic ID = 2 *The Good Doctor* spectators limit themselves to tweet that they are about to start a new episode of the TV series, the other two topics go deeper: in topic ID = 1, spectators express their impressions and perceptions about what they have just watched and in topic ID = 4 they speculate on the upcoming events in the narrative line. Turning to the topics related to the synchronous fruition, topic ID = 7 highlights that the hashtag used for the data collection trended during the airing time of the show along with other hashtags, implying that live-tweeting is a common practice among the fans of the series; topic IDs = 10 and 12 discuss the TV show schedules for the evenings in which the show airs and, lastly, topic ID = 17 publishes the most rated shows of the previous evening, in which *The Good Doctor* had resulted to be the most watched, suggesting

that most spectators enjoy watching it live. With regard to topics related to the asynchronous fruition, topic IDs = 15 and 18 both share one streaming platform through which Twitter users can watch the latest episode or the whole TV series. In topic ID = 19, the focus is on the sharing of YouTube videos made up of some lines collected from episodes of the show, in order to allow fans to watch (or re-watch) only scenes deemed salient.

The third most frequent category is formed by 6 topics that revolve around the fandoms of this drama, namely topic IDs = 5, 6, 8, 9, 11 and 16, totaling the 8.5% of the tweets. Among these, three topics aim the attention to Freddie Highmore's performance in playing Shaun, namely IDs =5, 9 and 11: if topic ID = 5 gathers tweets that praise him and his acting skills, topic IDs = 9 and 11 make emerge fans' engagement to indicate respectively their vote for him for the *People's Choice Awards* and to encourage other fans to do the same or their request for his nomination for the *Emmy Awards* in the category *Outstanding Best Lead Actor in Drama Series*. The other three fandom-related topics, i.e. IDs = 6, 8 and 16, outline petitions that fans submit: in topic ID = 6, the petition is directed towards the preservation of the presence of dr. Melendez' character in *The Good Doctor*, who dies at the end of season 3, while in topic IDs = 8 and 16, fans solicit for the renewal of the TV series for a new season.

In the fourth and last category, 4 topics that were considered noisy because they were not directly linked to *The Good Doctor* despite containing the hashtag *#thegooddoctor* were included, namely topic IDs = 3, 13, 14 and 20, amounting to the 4.8% of the tweets classified. In detail, topic IDs = 13, 14 and 20 exploit the popularity of the hashtag used for data collection, sometimes also combined with other trending hashtags, to promote respectively: a website that allows Twitter users to discover who has visited their profiles, a website that works only in the United States and a blog called *The Dahlinke2306 Daily*. Topic ID = 3 is slightly different from the others: this topic revolves around Twitter users explicitly indicating that they are sharing photos, but a close look at the shared photos yielded that they were not all related to the TV series. Additionally, the topic was classified as noisy because some other representative tweets were not directly interpretable or linked to the activity of sharing promotional contents connected to *The Good Doctor*.

Time Trends of Topics

Topics were also analyzed with regard to their evolution over time. Three time trends and respective features of the topics were hypothesized: con-

tinuous, discontinuous and peaking. Topics whose difference between the highest and the lowest point in the curve was lower than 0.3 in normalized frequency were credited as continuous; topics whose difference between the highest and the second-highest point was lower than 0.3 in normalized frequency were ascribed to the discontinuous trend, while topics whose difference between the highest and the second-highest point was higher than 0.3 in normalized frequency were considered to have a peak.

The time trends and their respective topics are reported in Table 5.

There are no topics that had a continuous time trend. Figure 3A and 3B display the evolution over time of the discontinuous and peaking topics respectively.

Considering the 13 topics that have a discontinuous evolution, what emerges by looking at Figure 3a is that the majority of their peaks and

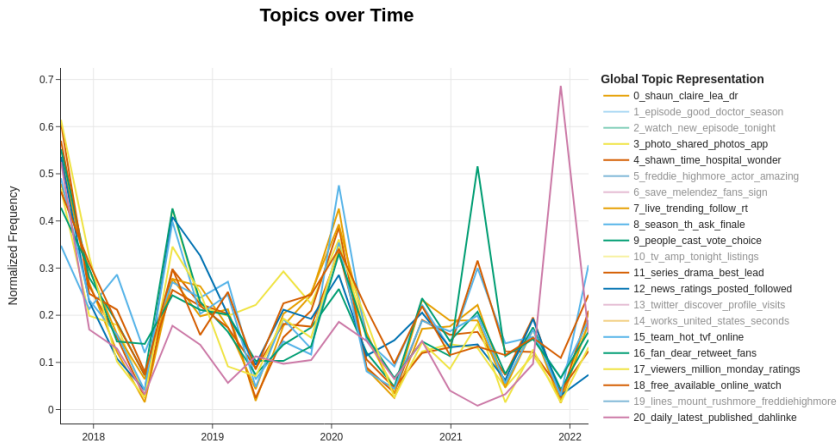| Time trend	Topic IDs	Total
Continuous	/	0
Discontinuous	0, 3, 4, 7, 8, 9, 11, 12, 15, 16, 17, 18, 20	13
Peaking	1, 2, 5, 6, 10, 13, 14, 19	8
TOTAL	/	21

TABLE 5

Time trends of topics.

FIGURE 3A

Time trend of the discontinuous topics, in normalized frequencies.
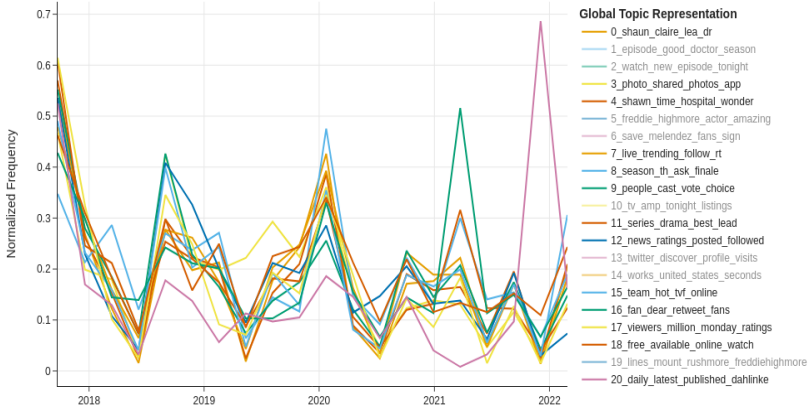
Topics over Time

FIGURE 3B

Time trend of the peaking topics, in normalized frequencies.

Season	Airing period	Christmas Hiatus	Peaks	Drops
1	September 25, 2017- March 26, 2018	December 4, 2017 – January 8, 2018	Last months, 2017	Approximately half 2018
2	September 24, 2018 – March 11, 2019	December 3, 2018 – January 14, 2019	Last months, 2018	Approximately half 2019
3	September 23, 2019 – March 30, 2020	December 2, 2019 – January 13, 2020	Approximately beginning 2020	Approximately half 2020
4	November 2, 2020 – June 7, 2021	November 30, 2020 – January 11, 2021	Last months, 2020	Between 2020 and 2021
			First months, 2021	Approximately half 2021
5	September 27, 2021 – May 16, 2022	November 22, 2021 – February 28, 2022	Last months, 2021	Right before the end 2021
			First months, 2022	/

TABLE 6

Seasons, airing periods and Christmas hiatuses of The Good Doctor, in parallel with peaks and drops of the discontinuous topics.

drops are somehow synchronized, the only two exceptions being the two noisy topic IDs = 3 and 20, respectively in yellow and purple. Accordingly, 7 common peaks and 6 consequent drops can be identified, which are reported in Table 6, in parallel with the corresponding airing season, airing period and Christmas hiatus for each one. By looking at Table 6, one can notice that the 7 peaks of the discontinuous topics occur approximately at the beginning of a new season of the TV series, in all the 5 seasons (5 peaks) and at the beginning of the TV series airing after the Christmas hiatus for the last 2 seasons (2 peaks). As a result, the 6 drops of these topics take place when the TV series is not airing, namely around the half of every year (4 drops) or during the last two Christmas hiatuses (2 drops).

Focusing on the 8 peaking topics displayed in Figure 3b, 5 of them show their peak during the last months of 2017, at the beginning of the airing time of *The Good Doctor*, namely IDs = 1, 2, 5, 10 and 19. Moreover, two topics peak little time afterwards, that is right before the end of 2017, i.e. IDs = 13 and 14. Only one topic peaks later in time, at the beginning of 2020: ID = 6. This can be explained because the topic is about the petition to save dr. Melendez and the character dies in S03E20, aired on March 20, 2020.

Sentiment Analysis Results

The results of the application of the RoBERTa sentiment analysis model on the *ASD dataset* are reported in Table 7. The number of tweets that were classified as positive is by far higher compared to those classified as neutral or negative, with the respective percentages being 59.5%, 29.3% and 11.2%.

In order to interpret these results, and to understand the main aspects towards which the sentiment of the tweets is addressed, we considered the 3 tweets with the highest score for each sentiment tag and their respective score (Table 8) and analyzed them qualitatively.

Considering the 3 positive tweets reported in Table 8, 2 tweets praise *The Good Doctor* directly for the representation of ASD (n. 1 and 3), and a

Sentiment	Number of tweets	Percentage
Positive	809	59.5%
Neutral	398	29.3%
Negative	153	11.2%
TOTAL	1,360	100.0%

Table 7
Sentiment analysis results.

personal experience with the condition is emphasized in both, either direct (n. 1) or familiar (n. 3), implying a reflection in Shaun's character by the audience. The positive sentiment of tweet n. 2, instead, is directed towards the show in general, even though the presence of one hashtag related to ASD draws the focus of the acclaim to the representation of this condition.

N.	Tweet	Sentiment tag and score
1	Omg Yes I'm #ActuallyAutistic and love this show can't wait to see season 5 in Australia so excited for Season 6 now 🌈😌😎 #TheGoodDoctor #series #TVShow #Autism #Neurodiversity #Australia https://t.co/Rldjbnq1F2	positive, 0.9896799
2	I am super excited for The Good Doctor to return September 24! #thegooddoctor #tvseries #doctor #show #autism #awesome #entertainment https://t.co/fk1ahMnxLG	positive, 0.9902643
3	Just watched #TheGoodDoctor and I love it! My son with #Autism also enjoyed it! Can't wait to see what's next!	positive, 0.990613
4	I had uploaded my 3rd "#Autistics in Fiction" video this time talking about Shaun Murphy from #TheGoodDoctor and #BenAffleck #autistic lead role in "The Accountant". ASPIE WITH ATTITUDE "Autistics In Fiction #3" E157 https://t.co/j7f1yprvET via @YouTube	neutral, 0.9157699
5	#TheGoodDoctor 1x12 "Islands Part Two" Promotional Photos & Synopsis https://t.co/Vj2EhIfPIt #FreddieHighmore #ShaunMurphy #Autism #GoodDoctor	neutral, 0.91611546
6	#TheGoodDoctor 1x07 "22 Steps" Promotional Photos, BTS & Synopsis https://t.co/yfsatSGu12 #FreddieHighmore #ShaunMurphy #Autism #GoodDoctor	neutral, 0.9311321
7	Found out recently #TheGoodDoctor got picked up for a 3rd season. *head-desk* Why?! This show is such a mockery of #ActuallyAutistic people. Makes me sick! https://t.co/Eu3V4USsvq	negative, 0.9443284
8	Asshole Doc needs to be FIRED. I'm sick of his arrogance and wrong judgment calls just bc he doesn't like Dr. Murphy #autism #thegooddoctor	negative, 0.9513848
9	#cancelthegooddoctor #TheGoodDoctor #ActuallyAutistic I have been waiting for fucking years for these assholes to apologize for working with autism speaks and putting on a stereotypical voice. You are hurting us.	negative, 0.9549853

TABLE 8
The 3 most representative tweets for each sentiment tag, their tag and their respective score.

As far as the 3 neutral tweets are concerned, 2 tweets do not address ASD directly, given that they are focused on the promotion of one episode of the medical drama (n. 4 and 5). Conversely, tweet n. 8 promotes the user's video from their YouTube channel in which they review ASD characters within audiovisual products.

Turning to negative tweets, if tweet n. 8 harshly criticizes a doctor's behavior towards Shaun, probably dr. Melendez or dr. Andrews, tweets n. 7 and 9 condemn how the condition is represented within the drama: tweet n. 7 defines *The Good Doctor* "a mockery of actually autistic people", while tweet n. 9 mainly blames the TV series for the display of stereotypes on ASD and for having collaborated with Autism Speaks, a non-profit autism-awareness organization that has been severely criticized by people with ASD, mainly because it considers the disorder as a disease to be cured.[7] Moreover, the author of the tweet asks for the cancellation of the drama through the hashtag *#cancelthegooddoctor* and wishes for an apology to people with ASD.

The fact that the sentiment of the *ASD dataset* is not consistently directed towards the representation of ASD within *The Good Doctor* must be considered a limitation of this research and could be overcome in future studies through an aspect-based sentiment analysis approach.[8]

Conclusions and Future Perspectives

This research aimed to investigate Twitter social discourses around *The Good Doctor* produced on Twitter, while simultaneously testing the adopted approach, based on NLP techniques. Pertaining to RQ1, whose purpose was to detect the main topics discussed around *The Good Doctor* by the audience and their evolution over time, the BERTopic topic modeling has showed that social discourses mainly revolve around the narrative lines of this medical drama and, more specifically, the storylines between the main characters. This finding validates previous research such as Castro-Mariño (2017)'s work, which observed, by coding social networks' posts on Spanish fiction programs, that the 62.6% of the comments were related to the narra-

---

[7]   For a detailed explanation of the controversy, please visit https://www.themarysue.com/the-autism-speaks-controversy-explained/ (last accessed 01-07-23).

[8]   An aspect-based sentiment analysis approach allows not only to extract the sentiment of a text, but also to determine towards which aspect (or category) the sentiment of the text is directed.

tion (Castro-Mariño 2017: 96). Almost as frequently as on narration, social discourses also focus on the fruition of this TV series: while some topics did not provide indications on the time, others suggested that the audience enjoys both watching the TV series live and live-tweeting to "[…] feel connected to a large online viewing audience" (Schirra et al. 2014: 2450) and on streaming platforms to follow one's own pattern of viewing. In fewer tweets, users have also demonstrated a high engagement as fans of the TV series, actively using social platforms to orient the accolades (topic IDs = 9 and 11) or submitting petitions to influence the narrative (ID = 6) or productive (IDs = 8 and 16) decisions around *The Good Doctor.*

The topics discussed by the audience were also displayed to have two main time trends: discontinuous and peaking. The majority of topics was discontinuous and followed the airing time of the TV series, whereas fewer topics presented a peak in the contemplated timespan, almost all during the first few months of airing. These findings suggest not only that the social discourses generated around *The Good Doctor* are strictly limited to the broadcast of the TV series and does not expand further in time, but also that they peaked above all at the beginning of its airing time.

As regards RQ2, aimed at detecting the sentiment of the audience towards the representation of ASD within *The Good Doctor* through the RoBERTa model, it has been shown that most tweets had a positive sentiment. This indicates that the representation of ASD within *The Good Doctor* has mostly been acclaimed, corroborating Stern and Barnes (2019)'s research, in which the authors observed that the representation of the disorder was positively received and judged as accurate by the surveyed people. The qualitative analysis of the *ASD dataset* has revealed, however, that the sentiment contained in the tweets was not consistently directed towards the representation of ASD and that a unanimous reception is far from being reached, neither among the general audience, nor among people with ASD, validating the ambiguity that had emerged from the literature (Hilsabeck 2022, Kluke 2020, Moore 2019, Stark 2020).

Methodologically, this first exploratory analysis has proved that NLP tools can be efficiently integrated into the field of reception studies of audiovisual serial products along with more traditional approaches. As a future perspective, we intend to expand the scope of this research by reproducing this fine-tuned techniques to investigate the reception of other US contemporary medical drama TV series, as well as shedding more in-depth light into the representation of ASD on specialized blogs using alternative methodologies.

BIBLIOGRAPHY

Amy (2022). "Hyperparameter Tuning For BERTopic Model In Python." *Grab N Go Info.* https://grabngoinfo.com/hyperparameter-tuning-for-bertopic-model-in-python/ (last accessed 01-07-23).

Andy, Anietie, Reno Kriz, Sharath Chandra Guntuku, Derry Tanti Wijaya and Chris Callison-Burch (2022). "Did that Happen? Predicting Social Media Posts that Are Indicative of What Happened in a Scene: A Case Study of a TV Show." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7209-7214.

Asbeek Brusse, Elsbeth D., Marieke L. Fransen and Edith G. Smit (2015). "Educational Storylines in Entertainment Television: Audience Reactions Toward Persuasive Strategies in Medical Dramas." *Journal of Health Communication* 20(4): 396-405. https://doi.org/10.1080/10810730.2014.965365.

Aulia, Silvia, Etika Rachmawati and Wawan Tarwana (2022). "The Use of the Movie "The Good Doctor" as Teaching Media in ESP." *Journal of English Education Program (JEEP)* 9(2): 1- 10. http://dx.doi.org/10.25157/(jeep).v9i2.8553.

Barbieri, Francesco, Jose Camacho-Collados, Leonardo Neves and Luis Espinosa-Anke (2020). "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." In *Findings of the Association for Computational Linguistics: EMNLP 2020,* edited by Trevor Cohn, Yulan He and Yang Liu, 1644-1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

Bouma, Gerlof (2009). "Normalized (Pointwise) Mutual Information in Collocation Extraction." In *Proceedings of GSCL* 30: 31-40.

Cambra-Badii, Irene and Joseph-E. Baños (2018). "Un médico con autismo en la televisión?: Enseñanzas de *The Good Doctor*." *Revista de Medicina y Cine* 14(4): 273-283.

Cambra-Badii, Irene, Ariadna Pinar and Joseph-E. Baños (2021). "*The Good Doctor* and Bioethical Principles: A Content Analysis." *Educación Médica* 22(2): 84-88. https://doi.org/10.1016/j.edumed.2019.12.006.

Castro-Mariño, Deborah (2017). "La audiencia social y el mundo narrativo de la ficción: análisis de los comentarios de los fans y community managers sobre la ficción televisiva y webseries españolas." *Palabra Clave* 21(1): 86-110. https://doi.org/10.5294/pacla.2018.21.1.5.

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)*, 4171-4186.

Dwiyanti, Cut Nadira (2022). *Language Disorder by the Main Character in "The Good Doctor" Series: Psycholinguistics Approach.* Master's Thesis. Batam: Putera Batam University.

Egger, Roman and Joanne Yu (2022). "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts." *Frontiers in Sociology* 7: 1-16. https://doi.org/10.3389/fsoc.2022.886498.

Grootendorst, Maarten (2022). "BERTopic: Neural Topic Modeling With a Class-Based TF-IDF Procedure." *arXiv:2203.05794v0571.* https://doi.org/10.48550/arXiv.2203.05794.

Hether, Heather J., Grace C. Huang, Vicki Beck, Sheila T. Murphy and Thomas W. Valente (2008). "Entertainment-Education in a Media-Saturated Environment: Examining the Impact of Single and Multiple Exposures to Breast Cancer Storylines on Two Popular Medical Dramas." *Journal of Health Communication* 13(8): 808-823. https://doi.org/10.1080/10810730802487471.

Hilsabeck, Burke (2022). "*The Good Doctor*: Images of Autism and Augmented Intelligence." In *Autism in Film and Television: On the Island*, edited by Murray Pomerance, and R. Barton Palmer, 71-82. New York: University of Texas Press. https://doi.org/10.7560/324912-005.

Kluke, Shauna (2020). *An Analysis of Social Competence Portrayed by the Autistic Character in the Television Series* The Good Doctor. Master's Thesis. Ottawa: University of Ottawa.

Koh, Sungran (2018). "An Analysis of English Vocabulary Based on Corpora from the Korean and American Versions of *The Good Doctor*." *STEM Journal* 19(4): 27-44. https://doi.org/10.16875/stem.2018.19.4.27.

Larasati, Ayu (2019). *Pragmatic Language Impairment in Autistic Savant Portrayed in* The Good Doctor *Film*. Thesis. Medan: University of Muhammadiyah Sumatera Utara.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv: Computation and language*. https://doi.org/10.48550/arXiv.1907.11692.

Mahdi Mohsin, Mohammed (2021). "Evidentiality in Doctors' Talk: A Discourse Analysis of "The Good Doctor" TV Series-Season One." *Adab Al-Kufa Journal* 48(2): 557-566. https://doi.org/10.36317/0826-013-048-066.

Moore, Allison (2019). "He's not Rain Man': Representations of the Sentimental Savant in ABC's *The Good Doctor*." *Journal of Popular Television* 7(3): 299-316. https://doi.org/10.1386/jptv_00003_1.

Molteni, Luca and Joaquin Ponce de Leon (2016). "Forecasting with Twitter Data: An Application to USA TV Series' Audience." *International Journal of Design & Nature and Ecodynamics* 11(3): 220-229. https://doi.org/10.2495/DNE-V11-N3-220-229.

Putra, Aditya Hayugraha Wahyu, M R Nababan and Sri Marmanto (2019). "Speech Acts Found in the Movie *The Good Doctor*." *International Journal of Multicultural and Multireligious Understanding* 6(5): 840-848.

Rokhim, Rizqi Fatkhur (2022). *Analysis of language disorder of the autistic savant character in* The Good Doctor *series movie.* Undergraduate thesis. Malang: Universitas Islam Negeri Maulana Malik Ibrahim.

Stark, Ian Scott (2020). *Representations of Autism in Popular Media: Rhetorical Disability in* The Good Doctor. Master's thesis. San Angelo: Angelo State University.

Schirra, Steven, Huan Sun and Frank Bentley (2014). "Together Alone: Motivations for Live-Tweeting a Television Series." In *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2441-2450. https://doi.org/10.1145/2556288.2557070.

Stern, Stephanie C. and Jennifer L. Barnes (2019). "Brief Report: Does Watching *The Good Doctor* Affect Knowledge of and Attitudes Toward Autism?." *Journal of Autism and Developmental Disorder* 49: 2581-2588. https://doi.org/10.1007/s10803-019-03911-7.

Tunstall, Lewis, Leandro von Werra and Thomas Wolf (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Sebastopol: O'Reilly Media.

twitter-RoBERTa-base model for sentiment analysis (2022). https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest (last accessed 01-07-23).

Tyasrinestu, Patricia and Priyatno Ardi (2020). "Idiomatic Expressions and their Indonesian Subtitles in *The Good Doctor* TV Series." *LLT Journal: A Journal on Language and Language Learning* 23(1): 37-57. https://doi.org/10.24071/llt.2020.230104/.

Ulatowski, Rachel (2022). "The Autism Speaks Controversy, Explained." *The Mary Sue*. https://www.themarysue.com/the-autism-speaks-controversy-explained/ (last accessed 01-07-23).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 6000-6010. https://doi.org/10.48550/arXiv.1706.03762.

ADOPTING NLP TECHNIQUES TO ANALYZE TWITTER SOCIAL DISCOURSES AROUND THE GOOD DOCTOR