The Relation between the Impact of Public Service Broadcaster's Talk Shows and Social Media Engagement

Sascha Hielkema

Master's thesis Applied Data Science, Utrecht University

Daily supervisor: Joris Veerbeek

First reader: Dr. Mirko Schaefer

Second reader: Dr. Max Kemman

July 2, 2021

Abstract

The rise of the digital world led to a shift from Public Service Broadcasters towards Public Service Media, where content is offered both offline and online. The Dutch Public Broadcaster NPO aims to produce content that has an impact on a diverse audience by making them feel connected to the society around them, increase their knowledge and touch their feelings. Therefore an impact score has been created by the NPO. On the other hand, social media websites became an important new part of the viewing experience of citizens. Now that viewers can express their opinions live during episodes, more information about the experiences of viewers can be measured and this could potentially be used as a metric to evaluate the content of television programmes. To gain more knowledge about the relation between online audience engagement and the impact of PSM television programmes, this research has been conducted. Impact scores of talk show episodes have been compared to the associated user engagement on Twitter, live during three NPO talk shows. Natural Language Processing tasks, such as sentiment analysis and text classification extracted sentiments and meanings from live posted tweets. These characteristics were added to a linear mixed model in which the impact score was predicted and talk show titles were added as mixed-effects. Results showed that the number of tweets and the percentage of negative tweets was related to a lower impact score, while the percentage of tweets about content indicates a higher impact score. The mixed-effects did explain some of the variance in the impact scores. Future research is needed to gain more knowledge about the relation between audience engagement and the impact of PSM content on individuals.

Keywords: Natural Language Processing, text classification, BERTje, RobBERT, logistic regression, Linear Mixed Model, Public Service Broadcaster, Public Service Media, Social media engagement, Twitter

The relation between the impact of Public Service Broadcaster's talk shows and social media engagement

Motivation and context

In the last decade, Public Service Broadcasters (PSBs) made a change towards Public Service Media. Their content distribution methods expanded from only traditional broadcast radio and television to digitally available content. This change, due to the rise of the internet, and especially social media, has affected public broadcasters in multiple ways (Van Dijck & Poell, 2014). For example, this change led to a revision of how public values are spread (Bennet, 2013). One problem arises from the conflict of interest between social networks with their commercial goal to attach consumers on the one hand, and the goal from public platforms to spread public values on the other hand (Van Es & Poell, 2020). PSM organizations should focus on their own goals and the spread of public values. Another change is the fact that their content can be created in new ways, for new audiences, in which active engagement of the audience is possible during programmes in live responses on social media (Kjus, 2009; Proulx & Shepatin, 2012; Van Es et al., 2016). This social media engagement could be used as a measure to reflect whether the content of television programmes meets up to the goals of a PSM organization.

In this thesis, the relation between the impact and social media engagement of the three most-watched daily talk shows from the Dutch Public Broadcaster (NPO) will be explored. The decision to focus on talk shows is based on two reasons. First of all, talk shows are known for showing controversial content and expressing different opinions in their programmes (Van Es et al., 2016). Secondly, there is a certain presumed live experience of viewers, that will possibly lead to more tweets about such programmes in real time (Van Es, 2016).

Previous research from Van Es and colleagues (2016), showed that the number of viewers and the number of tweets about television programmes from both commercial and public broadcasters were found as two distinct phenomena. To evaluate whether social media engagement could serve as a metrics for measuring television viewing experiences among citizens, further research about this relation is needed. Because this thesis is focusing on a PSM organization, it is less relevant to look at viewing numbers, but more meaningful to look at the distribution of public values, such as creating impactful content. As a result, different social media engagement features will be compared to impact scores of talk shows. The impact score is a measure created by the NPO that evaluates whether a television programme touches feelings, expands knowledge, and contributes to society culturally or socially according to a panel of representative Dutch citizens. Therefore, this research focuses on audience experience in a public service media setting, instead of commercial media.

To further analyse the relation between the impact of PSM television programmes and social media features, different tweet characteristics will be gathered using natural language processing techniques and machine learning algorithms. Specific tweet characteristics, such as the total number of tweets, sentiments, and information about the content of the tweets will be compared with the impact scores.

The goal of this thesis is to find what tweet characteristics are related to the impact score of PSM television programmes. The results will lead to more knowledge about the meaning of social media engagement in relation to PSM television programmes. More specifically, information about the relation between the impact of talk shows and social media engagement on Twitter will be gained. It will be an addition to the existing body of literature about PSM and fill the gap of information about its relation to social media engagement. Television makers could use social media engagement features in addition to the impact scores as a metric to find out whether a talk show episode meets up to their goals. Social media engagement could be measured directly after the broadcasted episodes, is less time consuming and easier to gather than the impact scores. Furthermore, when more knowledge about social media engagement is gained, engagement on social media could be promoted by the NPO to enhance user participation and to create more valuable discussions online.

Based on the literature, a theoretical framework will be presented in which a data science question from the media domain is embedded. This leads to the research question of how social media engagement and the impact of talk shows from the NPO, are related to each other. This question will be answered and discussed throughout the rest of this paper. The acquired data and used methods will be described. The results will explain the relation between the variables, and whether there is a significant relation between tweet characteristics and the impact of talk shows. Finally, a conclusion, implications, and suggestions for further research will be discussed.

Literature overview

Social media platforms, such as Twitter, Facebook, and Instagram have become an integrated part of today's society. Individuals in society can easily share their thoughts, opinions, and experiences online (Page, 2017). The increasing importance of the digital world changed the media landscape for Public Service Broadcasters that need to serve the democratic, social, and cultural needs of society. (Van Es & Poell, 2020). Consequently, public broadcasters changed their traditional ways of distributing public values to public service media (PSM) (Bennet, 2013). Next to linear television, PSM offers social and interactive content and services that guarantee quality, diversity, and democracy in media. Compared to the past, public broadcasters often redefined the public interest, diverse, reliable, and quality information due to the changing society and changing media landscape (Bardoel & Brants; 2003).

The rise of digital media created challenges and opportunities about how PSM should shape public values. For example, there are new ways to measure audience engagement on social media, which can be used to observe the effects of PSM content on individuals. Besides, the spread of public values, such as independence, trustworthiness, pluriform, diversity, engagement, and impact, is extended to multiple channels that can reach the audience (Van Es & Poell 2020). On the contrary, commercial social media platforms could threaten the spread of public values. Several of these changes, their consequences, and the chances for the NPO will be discussed in this chapter.

NPO

This thesis is focused on the Dutch Public Broadcaster (NPO). The NPO is an umbrella organization governed by the Dutch Foundation where multiple independent broadcasting organizations belong to. Last century, these organizations were all reflecting a societal, religious, and ideological movement in the Netherlands, known as pillars, such as Catholicism, Protestantism, Socialism, and Liberalism. Essentially, the exposure of other pillars on television to individuals resulted in pluriform content and finally the depillarization of the Dutch society (Engelbert & Awad, 2014). The different independent broadcasters within the NPO organization still exist nowadays. Each of them has their own identity and goals, but the structure seems less relevant to reflect the whole diverse society these days.

In a letter from the parliament written by Arie Slob, head of the Ministry of Primary and Secondary Education and Media (2019), the government states that changes are needed in the system of the NPO to fully serve society as a PSM. The NPO has the ultimate responsibility for the content produced by the separate broadcasters. At this moment, the separate broadcasters mostly determine how the content is offered and showed to the audience. One problem that occurs due to this regulation is that the distribution of content does not completely fit the needs of the audience. Due to a continuation of changes within society and a more diverse audience, not all societal pluriformity is reflected within the broadcasters. The NPO should be aware of these societal changes and improve their collaboration with the separate broadcasters to fit their content to the needs of all citizens.

Another important change is the rise of possibilities for citizens to unite online. On social media, citizens show their engagement in communities. The NPO should be up-to-date about such societal changes and adapt their policy to these changes. While in the past, a membership for a specific broadcaster was used to reflect the engagement of viewers, nowadays, engagement of viewers and their interaction can be measured in other, more relevant ways, such as social media engagement.

Social media engagement

Nowadays, public broadcasters are producing content in new ways, for new and more active audiences (Kjus, 2009). Compared to the past, a new personal and interactive relation between the viewer and PSM has been created. The use of digital platforms can enhance forms of public participation and reach new types of audiences. Social media websites became an important new part of the viewing experience. Besides, television watching behaviours have changed as well. Currently, multiple devices, such as laptops, tablets, and smartphones are used to watch television on demand (Gillan, 2011). Online discussions take place about the topics covered in television programmes, in which especially young people are actively engaging (Bober, 2014). In the past, the broadcast spectrum was limited, with a limited viewer choice and a passive role for the citizen. Whereas in the digital world, platforms are offering an abundance of offers for a more interactive consumer in a niche community (Bennett, 2013).

Van Dijck and Poell (2014) highlight the importance of PSM in using social media to engage younger audiences in public television. Young audiences are shifting from linear television to videos on demand (Schwarz, 2016). No involvement in social network sites would lead to a loss of this younger generation of viewers. Besides, it would lead to a loss of young makers and programme producers, which should be avoided.

There are several ways to measure audience engagement. An article from Bober (2014) described how a connection is made between reality TV, sports broadcasts, and their viewer engagement on social media. Tweets and Facebook messages were used to investigate the interaction of viewers with programmes. In this article, different Twitter features were extracted that represented viewer engagement, such as the volume of tweets before, during, and after broadcasts based on programme-specific hashtags. For both television genres, the most activity was found during the end of the episodes for users that watched the episodes linear.

A large body of literature exists on the relation between social media engagement and big sport or cultural events. Research from Highfield and colleagues (2012) was focused on tweets about the Eurovision Song Contest, an international television programme broadcasted over Europe. In that research, Twitter is seen as a technology that measures fandom and the feelings of belonging to an audience for such events. Individuals in these audiences were connected to each other on Twitter. A more recent research from Hagen and Stauff (2021) showed how controversies in sport's live events intensify discussions on Twitter

To evaluate the effects of the content of PSM television programmes on individuals, NPO could use social media engagement features as a metric. Information from tweets can be used to reflect what viewers think of the content presented to them. According to the goals from the NPO for the upcoming five years, the organization wants to promote user engagement, interaction and build a profound connection with viewers online. Therefore, different channels are created on Facebook, Instagram, YouTube, and Twitter, to post messages about television programmes, encourage viewers to react, and make them feel connected to the content (NPO, 2020).

Public values

Next to new opportunities, such as promoting audience engagement, the rise of new media also created challenges for PSM. An article from van Dijck and Poell (2014) describes how the shift to the online world has led to a change in the meaning of publicness on institutional, professional, and content-level. Social networking services give the audience access to the online production and distribution of audio-visual and textual context. The commercialization of social networks threatens to compromise public values. Citizens are often seen as consumers through social media platforms that use their data for commercial purposes (Murdock, 2018; Van Es & Poell, 2020). This ecosystem of social media leaves very little to no space for non-commercial non-profit, public platforms. While Netflix and YouTube are known for offering personalized recommendations to users, such recommendation algorithms are in tension with the public values of diversity and universality in PSMs (Van Es, 2017). PSM should aim to achieve other objectives.

In the past, a similar challenge did occur between PSM and commercial competitors. Due to the rise of commercial broadcasters, a lot of viewers transferred from the public broadcaster to watch programmes from the commercial broadcasters (Bakker & Scholten, 2003; Picard, 2002; Meijer, 2005). On the other side, commercial broadcasters complained about the unfair competitive advantages of public service broadcasting programmes, which seemed to lack public values (Syvertsen, 2003; Van Dijck & Poell, 2014; Benson et al., 2017; Garcia-Martinez & Nguyen, 2012; Sjøvaag et al., 2016). However, unlike the commercial competitors, public service media should really focus on showing public values in their content. The Mediawet (2008), a Dutch law, states the role of the NPO within society, to administer public service broadcasting. This law assures that the NPO focuses on its public role instead of commercial motives.

Impact

The strategies of PSM have to contain certain public values, which should be showed to the audience. To connect the audience to society, NPO wants to create impactful content (NPO, 2020). The role of impact has already been studied in the past. Meijer (2005) highlights that the quality of TV programmes consists of the content of a programme, the involvement of the audience, and the impact of programmes. Impact is described as high quality content containing information based on the truth and involving people with diverse backgrounds. In this context, the viewer is seen as a citizen, consumer, and enjoyer of the programmes. Van Es and Poell (2014) also describe how impact is seen as a public value. In this framework, impact means that the content reflects social cohesion, the development of users, cultural values, and quality of democracy and society. In addition, the European Broadcasting Union (2012) described six core values that PSM needs to account for. Universality, excellence, diversity, and accountability all contribute to the impact of programmes, since these values aim for relevant, high quality, diverse and accurate content.

The NPO aims to inform, inspire and entertain citizens. Besides, the broadcaster wants to connect individuals in the Dutch society online and offline by producing meaningful television programmes. Accordingly, the NPO is creating media that has an impact on society. In this context, impact means that a television programme evokes certain feelings, improves knowledge, and has a perceived contribution to society according to viewers. An impact score measurement has therefore been developed to measure viewer's engagement and the impact of programmes on individuals in society. The impact is measured among a panel of representative Dutch citizens from the Media Appreciation Panel from the GfK (Growth from Knowledge), who watched the television programme. Participants from this panel answer multiple questions on a Likert scale and indicate whether a programme touches their feelings, makes them feel connected to others, expands their knowledge, considers the content as useful, and if they feel like the program contributes socially or culturally to society. Thus, the experience of the audience has a central role. All these five subcategories will be measured and each will result in a score between one and hundred. The mean of the five results is equivalent to the overall impact score. This method is validated by professor Bijmolt from the University of Groningen in 2019 (NPO, 2021).

Relation between online user engagement and impact

For the upcoming five years (2022-2026), the NPO created a policy plan, in which the four main goals of how the public broadcaster wants to be valuable are described. The first goal is to offer diverse and impactful content to the audience, in which high quality and relevance are present. This goal will have different subgoals per genre. The second goal is to provide next to linear television, corresponding on demand and online content on different channels, to increase viewer engagement. On online channels, interaction, participation, and in-depth content will be promoted, so that the content is in accordance with the needs of the audience in the digital society. The third goal is to connect individuals to society, which is included in the impact score of TV programmes. The last goal, which will not be further discussed in this paper is that the NPO is accessible and recognizable (NPO, 2020).

While in the past, the measurement for audience engagement was evaluated by the memberships per broadcaster, nowadays new measures are needed to reflect the effect of television programmes content on individuals. The NPO is responsible to offer content that fits the needs of their audience. Therefore, audience engagement could potentially be a metric to evaluate the content. Audience engagement in this thesis will be measured by social media engagement of viewers on Twitter in this research. The relation between an important goal from the NPO, creating impactful programmes to connect the audience to society, will be compared to social media engagement on Twitter.

Research from Van Es and colleagues (2016) already showed that there are differences in the number of tweets and the number of viewers per television programme. Their article suggested that other features should be used to continue the exploration of the relation between social media and viewing engagement behaviours of television audiences. They also explain that political talk shows have more controversial content. Besides, talk shows have a higher live experience among viewers, which would lead to more reactions (Van Es, 2016). To further analyze possible associations between social media engagement and viewing behaviours, in the context of PSM organization, the impact of television programmes on individuals and the corresponding Twitter characteristics have been measured.

To evaluate the experience of viewers from public service media talk shows, social media engagement seems a meaningful measure that contains information about the level of interaction and engagement of viewers. This thesis will add new information to the existing body of literature about social media engagement and public service television programmes. Because PSM values the spread of public values instead of high viewing numbers, there will be focused on the sentiments of tweets and the content of the tweet, to find out whether the tweet is valuable in relation to the episode. The goal is to find what tweet characteristics indicate a more impactful episode. This information will fill the existing gap in the literature about the relation between social media engagement and PSM content. Besides, information about the relation between these concepts would lead to more knowledge and better insights into the current policy from the independent broadcasters, and the NPO in general. In this context, it would be relevant to make a connection between the impact score of talk shows from the NPO and the engagement of the audience online on Twitter. This leads to the following research question: What types of tweet characteristics are associated with the impact of television programmes on society?

When a programme has a high score on impact, this could mean that there were more positive, negative, and valuable tweets during the programme. While a programme that scores lower on impact could lead to less positive, negative, or valuable tweets. If Twitter features, such as sentiments of tweets or the specific content of tweets are significantly associated with the impact of television programmes, programme makers could focus on promoting such discussions online or in their content of tv programmes to enhance user engagement and create more valuable discussions online.

The rest of this thesis will focus on the analysis of this association, and compare the impact scores from talk shows with different Twitter features, representing social media engagement. The engagement features have been extracted from Twitter, based on hashtags, like described in the article from Bober in 2014. The results show what type of tweets are associated with the impact of talk shows episodes. Sentiments could show whether viewers think positively about the show or that they tweet more negatively about it. The content of tweets will indicate whether people talk more about the content or other topics in the show. A linear mixed-effects model will be interpreted and finally, the results will be interpreted, so that insight will be gained about the relation between the impact of PSM talk shows and social media engagement.

Data

To find potential relations between the impact of talk shows and the associated audience engagement characteristics on Twitter, data from two different sources are combined and analysed in this paper. Information about the three most-watched daily talk shows from the NPO: *M*, *Op1*, and *de Vooravond* has been used to find this relation. I obtained access to a dataset containing information about television programmes through the NPO. The second dataset contains Twitter data, in which tweets about the talk shows are collected. Together, these datasets are combined into a final dataset aggregated by episodes. More specifics about these two datasets will be further described in this section.

Talk show data

The first dataset contains general information about the talk show episodes, such as the length, the date, and channel owners. In addition, viewing numbers, impact scores features, and the total impact scores of talk shows can be found in this dataset. The information was accessible via Google Cloud Platform and extracted by filtering on the titles of interests using SQL via BigQuery. The impact scores are measured by a questionnaire from the GfK. A panel of 9,000 Dutch citizens older than 13 years answered several questions about the programmes they have watched on a Likert scale. The final impact score is the weighted average of the features that measure whether a program touches one's feelings, connects individuals with the world around them, expands their knowledge, shows useful information, and whether the episode has a positive contribution to society. Scores of these separate features were provided. The average score was calculated in SQL. All talk show episodes which had less than 30 respondents to one of the impact score features, were excluded from the final dataset because this could lead to generalization problems and possible biases in this research. Except for the days on which there was no talk show episode, there was no data missing. General information about the seasons and total included and excluded episodes of the three talk shows can be found in table 1.

Table 1

Information about the talk shows

Title	Seasons	Start date	End date	Included episodes[a]	Excluded episodes[b]	Impact score
Dit is M	2	2020/06/01	2020/07/03	18	7	68.2
		2020/10/26	2020/02/12	56	14	69.2
Op1	1	2020/01/06	2021/02/22	190	24	72.1
Vooravond	2	2020/08/31	2020/10/23	39	1	70.0
		2021/02/15	2021/04/19	43	1	72.5
Total				343	48	70.4

Note. These talk shows are broadcasted during the seasons on a daily basis, excluding weekends.
[a] These impact score features were included because it had more than n=30 respondents
[b] These impact score features were excluded because it had less than n=30 respondents

Twitter data

Twitter data has been extracted by OBI4WAN, a Dutch social media monitoring website. An educational license has been obtained via the University of Utrecht, which allows to scrape up to 20.000 tweets at a time based on hashtags over the past year. This process has been repeated to get the total amount of tweets for the talk show episodes included in the first dataset. Tweets were collected with the programme-specific hashtags *#DitIsM, #Op1,* and *#Vooravond* between the 1st of June 2020 and the 20th of April 2021. All the tweets per talk show were extracted separately, and the name of the talk show, about which the tweets were posted, was added to the data so that there was no confusion as to which talk show the tweets belong to. Because this research is focused on the live experience of viewers on social media during the talk shows, only the live tweets that were posted by viewers during the show or at least until one hour after the show were kept in the dataset. In addition, tweets from the talk show profiles themselves were deleted, because those tweets did not contain information about viewer engagement. Finally, the tweets were combined, resulting in a final dataset of 216,970 tweets. The information about the tweets was aggregated per day and combined with the episode information. This leads to the following number of tweets per talk show, as can be found in table 2.

Table 2

Total tweets per talk show

Title	Seasons	Start date	End date	Number of tweets
M	2	2020/06/01	2020/07/03	15,342
		2020/10/26	2020/02/12	26,201
Op1	1	2020/01/06	2021/02/22	131,804
Vooravond	2	2020/08/31	2020/10/23	12,477
		2021/02/15	2021/04/19	31,146

Outliers

Several outliers have been detected that were three standard deviations (Z>3) from the mean (Osborne & Overbay, 2004). Two episodes from *M* and one episode from *Vooravond* had impact scores around 58 and 59. There was some variation in the number of tweets per episode. Several episodes lead to a very high number of tweets compared to the average tweet count per episode. However, all these values that could be considered as outliers are kept in the dataset. These values are meaningful and especially of interest for the analysis because the difference between low and high numbers of tweets per episode is relevant to find the relation between the impact scores, and therefore they have been included in the dataset.

Data preparation

To prepare the Twitter dataset for feature extracting and content classification, the data got cleaned first. To clean the corpus, regular expressions were used to erase all irrelevant symbols, punctuations, emoticons, URL links, and special characters out of the tweets so that a new corpus without irrelevant characters was available for further natural language processing, which will be further described in the method section.

Ethical considerations

In this research, data from different sources have been collected, which cause different ethical considerations. The data from the NPO is based on programme specifics and contains information about the impact that television programme episodes have on individuals. The respondents were voluntarily participating in the research and were aware that their answers were measured and used for media research by the Media Appreciation Panel. No personal information that could lead back to an individual was provided in the dataset. Therefore, the assumption holds that no ethical issues have been raised in this part of the data.

Nevertheless, the second dataset does involve more personal information from Twitter users. Besides, these users have not been notified about the use of their tweets in this research. However, only tweets from users with a public profile have been extracted, which is in accordance with Twitter's policy (Twitter, 2021). Therefore, it is assumed that Twitter data meets ethical guidelines. In the dataset, only the usernames of viewers and their tweets can be found. After the removal of tweets from the NPO programmes themselves, all the usernames have been deleted from the dataset. Only usernames in retweets are left in the non-cleaned corpora. No other personal information about the users is collected. Ideally, the users should have been asked for consent to use their tweets, to reduce ethical concerns. However, this is not possible for this research. Besides, Twitter's terms of service highlight that the information of users can be accessed by third parties. Twitter users that accept these terms, automatically give consent for the use of their data, which meets legal guidelines (Williams et al., 2017).

Methods

After exploring the datasets, different methodological approaches were used to explore the relationship between Twitter characteristics and the impact scores of three talk shows broadcasted by the NPO. Several natural language processing tasks (NLP) have been performed using Python 3.8.10 to extract different characteristics from Tweets. This has been done by using machine learning techniques, such as sentiment analysis and text classification. These features were added to a linear mixed-effect model (LME) using RStudio version 1.3.959 to discover their contribution to the impact score of the talk shows. An elaborate overview of the methods used in this research will be discussed in this section.

Sentiment analysis

A sentiment analysis has been performed to classify the cleaned tweets, that were posted live during the talk shows, as either positive or negative. Lately, sentiment analysis has earned recognition in research fields (Perk & Paroubek, 2010). When subjective feelings of individuals can be detected, more knowledge about human behaviour will be gained, which brings benefits to organizations (Vyas & Uma, 2019; Eke et al., 2021). The NPO could use information about the most prominent sentiments in tweets, to see whether they relate to impact scores of television programme episodes.

Sentiments can be extracted from texts using NLP tasks. The Dutch pre-trained model BERTje (de Vries et al., 2019) from the Hugging Face Transformer library has been used to perform a sentiment analysis on tweets about the selected talk shows episodes from the NPO. This model is a variant of BERT, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). BERT is a Transformer language model with state-of-the-art results for NLP tasks. The model is a pre-trained on Wikipedia texts, that learns information within and between sentences. Devlin and colleagues (2019) describe that “BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" (p. 4171). The model can learn to embed each word on the surrounding words and find semantic coherence between sentences.

Because this research is based on Dutch tweets, the Dutch model BERTje has been used, which is a pre-trained model able to perform several language tasks, such as sentiment analysis. BERTje is trained on a Dutch Book Review Dataset with 110.000 positive and negative labeled Dutch reviews. The model performs the NLP task of labeling sentiments to Dutch texts with an accuracy score of 93.8% (Van der Burgh & Verberne, 2019). Even though the tweet corpora used in this research varies from the pre-trained reviews, there is assumed certain generalizability of this model to tweets, due to the informality of the texts. However, the accuracy is expected to be lower in this other context. The model was loaded and the tweets were classified by the model for sequence classification. The total number of tweets and the number of positive and negative sentiments during an episode were aggregated and added to a dataset containing information about the episodes and impact scores.

To evaluate the model performance, a sample of 250 tweets has been manually checked on the accuracy of the classified labels. To assure that the sentiments completely reflect the content of the tweets, some tweets were checked if they better belonged to a neutral label. Accuracy has been calculated by the true positive and true negative tweets divided by the total number of classified sentiments. The accuracy of the model was 83%. However, this is not an accurate reflection, because the tweets are limited to 280 characters, and not all those tweets will present a clear sentiment. In table 3a an evaluation matrix can be found that reflects the performance of the classification tasks when the actual neutral tweets were left out. In table 3b a confusion matrix can be found, which shows that a neutral class would be a good addition to the model.

Table 3a

Evaluation matrix excluding neutral label

Label	Precision	Recall	F1-measure
Positive	.70	.74	.72
Negative	.89	.86	.87

Table 3b

Confusion Matrix

n = 250	Predicted positive	Predicted negative	Total
Actual positive	32	11	43
Actual negative	14	89	103
Actual neutral	26	78	104
Total	72	178	250

Text classification

To get more insight into the content of the tweets posted by viewers, a text classification task has been trained on manually labeled tweets. In addition to the sentiments of tweets, the content of the tweets was classified according to three different labels of interest. It would be interesting to find out if people who experience high levels of impact from talk show episodes, also post more valuable tweets about these episodes. Among all the tweets, a lot of posts were found containing either a reaction to the *content* of the talk show, *person(s)* in the talk show, and *other* subjects outside of the talk show. The explanation of the classification guidelines can be found in table 4. A total of 1,326 tweets has been labeled. From these tweets, all the duplicates, known as retweets, were deleted, resulting in a dataset of 1,002 tweets. The retweets were removed to prevent potential biases in the evaluation of the accuracy of the text classification models.

Table 4

Classification scheme: labels assigned to the tweets

Label	Description
Content	Tweets with opinions or shared information about the content of talk shows can vary between the music played in the show, the political messages spread in the show, or any other issue that is discussed during the talk show episode
Person	Tweets with opinions about the talk show hosts or the guests in the show. If there is an opinion about a person who is not part of the show, it is seen as content
Other	All other tweets that are neither containing opinions and information about the content nor persons in the talk show. These tweets can be funny comments, or more superficial comments on the content, with no opinions or relevant information. All tweets that were doubted or containing too many errors were assigned as other too. This label also contains tweets with only hashtags or @-mentions, other languages, and tweets about external videos

Different classification models have been evaluated on their performance to apply the right tags to the labeled content. Consequently, these machine learning algorithms were compared to each other to find the best classification model. A simple logistic regression Bags of Words model has been selected as the baseline model. Three other models were the TF-IDF regression model and the two Dutch models BERTje and RobBERT. From the 1,002 tweets, 430 were labeled as *content*, 189 tweets were labeled as *person* and 383 were labeled as *other*. The dataset was further divided into an 80% training set, a 10% test set, and a 10% validation set for all models. Even though the validation set is only required for BERT models, it is used for the logistic regression models as well to assure a fair comparison between the different classifiers. Due to the imbalance between the labels, as can be seen in table 5, the classes in all models were mapped to weight values, which makes the distribution of labels more equal, and therefore lead to a more accurate performance of the classification models.

Table 5

Train, test, validation set distribution

	Training set	Test set	Validation set
Content	348 (43.5%)	43 (43.0%)	39 (38.6%)
Person	148 (18.5%)	19 (19.0%)	22 (21.8%)
Other	304 (38.0%)	38 (38.0%)	40 (39.6%)
	800	100	101

Text classification model training

Before the BOW-model was trained, the corpus was tokenized. After that, the training set was used to create a vocabulary. Each sentence was vectorized, and the occurrence of the words in the vocabulary was counted. The test data was used to evaluate the performance of the BOW model. An overall accuracy of 68% was found for this model.

The second logistic regression model that has been used for text classification in this thesis, extracted frequency-inverse document frequency (TF-IDF), which generates vectors of weight within a specific document and adds more weight to it when the word occurs less frequent in the total corpus. So that the weights are adjusted for more or less frequently used words (Huang et al., 2007; Wang et al., 2017). The test set showed an accuracy of 58% for this model.

In third place, the Dutch model BERTje has been used to classify the text. This model has already been described in this thesis for the classification of sentiments on tweets. Compared to the other two models, BERT is a smarter model that can learn semantic information, such as coherence within and between sentences (Devlin et al., 2019). However, the general Dutch-based model is pre-trained on several corpora, such as books, news, references, news websites, and Wikipedia, representing 12GB of text (de Vries et al., 2019). The model is fine-tuned on the training set for five epochs. Longer training on the validation data decreased the performance of the model. The hyperparameters were l_r = 5e-6, and a batch size of 32. The test set achieved an accuracy of 60%.

Finally, the second version of the Dutch model RobBERT was loaded to classify labels to the text. This model is based on roBERTa, an optimized BERT model. RobBERT outperforms several other models on smaller datasets. The model is trained on a Dutch corpus and a Dutch tokenizer was used to tokenize the texts. The model is trained on a 39GB large Dutch corpus from OSCAR (Ortiz Suarez et al., 2019), which is much larger than the corpus used for BERTje. The model is fine-tuned on the tweet corpus similar as described for the BERTje model with *lr* = 5e-6, and batchsize 32. Again, the validation set evaluated after each epoch whether the performance improved. After four epochs the model stopped running. The test set showed an overall accuracy of 65% which ultimately motivates the choice to use the robBERT classification model to classify the tweet corpus of this research. All the results from the models can be found in table 6. After the model had been compiled, trained, and evaluated, the unlabeled tweets were classified based on the trained model

Table 6

Model performance

Model		Precision	Recall	F1-measure
Logistic Regression Bags of words	Content	.74	.60	.67
	Person	.57	.42	.48
	Other	.57	.76	.65
	Macro average	.63	.60	.60
	Weighted average	.64	.63	.63
Logistic Regression TF-IDF	Content	.63	.67	.65
	Person	.54	.37	.44
	Other	.54	.58	.56
	Macro average	.57	.54	.55
	Weighted average	.58	.58	.57

Model		Precision	Recall	F1-measure
BERTje	Content	.59	.67	.63
	Person	.62	.42	**.50**
	Other	.61	.61	.61
	Macro average	.60	.57	.58
	Weighted average	.60	.60	.60
RobBERT	Content	.81	.67	**.73**
	Person	.41	.63	**.50**
	Other	.69	.63	**.66**
	Macro average	.64	.65	**.63**
	Weighted average	.69	.65	**.66**

Note. The bold scores represent the highest accuracy in this comparison

Linear Mixed-Effects Model

To explore the relationship between the twitter characteristics and the impact scores of talk shows, a linear mixed-effect regression model has been built in RStudio. This model incorporates both fixed- and random-effects to take into account different baselines for specific categorical variables of interest (Koerner & Zhang, 2017). The number of tweets was normalized to z-scores and added to the linear mixed model as fixed effects. Furthermore, the values describing the percentage of positive and negative tweets, and the values that represent the content of the tweets were added as percentages to the model as fixed effects. Talk show titles were added to the model as random effects. These random effects create different intercepts per talk show title. After that the assumptions of linearity, normality, multicollinearity, and homoscedasticity were checked, the model was built using R's package lmer4.

In the field of linear mixed models, there is an ongoing debate about the inclusion and interpretation of p-values of the fixed effects in mixed models. To get *t-values* and *p-values,* a normal approximation has been used, in which an infinite degrees of freedom is assumed so that the t-values can be treated as z-values (Williams, 2018). In previous research, these values have been used to evaluate the significance of the mixed models (Luke, 2017). In this paper, the significance of the parameters will be interpreted as well. The model fit has been reported by R’s package MuMIn.

Results

The goal of this research was to find what Twitter characteristics are associated with the impact of talk shows from the NPO. As explained in the methods section, different natural language processing tasks were employed to extract characteristics from tweets. After that, a linear mixed model analysis was performed, to find out which these variables serve as independent predictors of the impact scores. The results of these analyses will be presented in this section.

Distribution of twitter characteristics

The different sentiments and subjects of tweets were extracted, aggregated to episode level, and divided by the total number of tweets per episode. This resulted in percentages, of which an overview can be found in table 7.

Table 7

Tweet characteristics distribution

Title	Number of tweets	Positive	Negative	Content	Person	Other	Average impact
M	41,543	12,615 (30.4%)	28,928 (69.6%)	21,379 (51.5%)	7213 (17.4%)	12,951 (31.2%)	69.0
Op1	131,804	40,530 (30.8%)	91,274 (69.2%)	67,051 (50.9%)	22,703 (17.2%)	42,050 (31.9%)	72.1
VA[1]	43,623	13,942 (32.0%))	29,681 (68.0%)	20,074 (46.0%)	9,391 (21.5%)	14,158 (32.5%)	71.7

Note. [1] Vooravond

Multicollinearity

Almost all of the independent variables were weakly correlated, which means that these variables independently contribute to the overall effect of the impact score. One of the independent variables, the percentage of negative tweets, also contained information about the percentage of positive tweets. Therefore, only one of these variables is used in the final model.

However, two independent variables had a very high correlation, the percentage of tweets about content and the percentage of tweets about other topics. This is caused by the fact that the weighted percentages of content, person, and other tweets are together representing a score of one. Due to the small contribution of tweets about persons, the percentages of content- and other tweets, are highly correlated according to Pearson *r*(341) = .88, *p* <.01. This motivated the choice to exclude the percentage of other tweets. In addition, the variance inflation factors (VIF) were checked and resulted to be all below four, causing no concerns for multicollinearity (Miles & Shevlin, 2001). The underlying correlations and VIF scores can be found in table 8.

Table 8

Correlation matrix for independent variables

	Number of tweets	Percentage negative	Percentage content	Percentage persons	Percentage other	VIF
Number of tweets		r = .20	r = .30	r = -.09	r = -.37	1.191
Percentage negative			r = .02	r = .02	r = -.04	1.045
Percentage content				r = -.75	r = -.83	2.437
Percentage persons					r = .25	2.246
Percentage other						2.673

Mixed Model

All the independent variables, except the percentage of other tweets, were added to the linear mixed-effect regression model as fixed effects. Because there were differences found in the average impact scores per talk show title, as can be found in table 7, the data should be nested by the talk show titles to allow differences between the mixed model intercepts.

Consequently, talk show titles were added as random effects to the model. The model was fitted using maximum likelihood (RL). The results of the mixed model can be found in tables 9 and 10.

Table 9

Summary of random effects Mixed-effects model

	Variance	Standard deviation
Talk show title intercept	1.685	1.298
Residual	10.678	3.286

Table 10

Summary of fixed intercepts Mixed-effects Model

	B	SE B	t	p
Intercept**	70.041	2.423	28.902	.000
Number of tweets*	-.434	.194	-2.232	.026
Percentage negative**	-6.880	2.314	-2.973	.003
Percentage content**	10.156	2.287	4.441	.000
Percentage person	4.640	3.734	1.243	.214

Note. $R^2m = .099$, $R^2c = .222$
* Significant results at the .05 level;
** Significant results at the .01 level

The variance around the random intercepts per talk show title was 1.685 and the residual variance was 10.678. This indicates that there is more variance in the residuals. However, the random effects per title do explain some of the differences in the impact scores, which means that there is more variance within talk show titles than between the different titles. The intercept coefficient in table 9 lies around the average of the talk show episodes b= 70.041, t(336)= 28.902, $p < .001$. Because this mixed model plotted different intercepts per talk show title, this number varies between the titles.

The normalized number of tweets b= -.434, t= -2.232, p = .026 is significantly negatively correlated to the impact scores. This variable was normalized, which means that an increase of one standard deviation in the number of tweets decreases the impact score by .434. Besides, the percentage of negative tweets b= -6.880, t= -2.973, p = .003 was also negatively correlated to the impact score. Because a score of one represents a hundred percent, the beta coefficient can be interpreted as follows: if the percentage of negative tweets in a talk show episode increases with one percent, the impact score decreases with -.068 standard deviations. A higher percentage of negative tweets per episode results in a lower final impact score on average. On the other hand, the percentage of tweets about the content of the talk show positively correlates with the impact scores b= 10.156, t= 4.441, p < .001. If the percentage of content full tweets increases by one percent, the impact score will increase with .102 standard deviation. Finally, the percentage of tweets about persons in the talk show episode did not show a significant result b= 4.640, t= 1.243, p = .214. The beta coefficient is very small but is found to positively correlate to the impact score. However, due to the high standard error, there is a lot of uncertainty about this coefficient, so no conclusions about this relation can be drawn.

The variance explained by the model is normally reported by the R-squared value. However, in mixed models, this value is often not reported. It is difficult to include this score, due to the random effects account for explaining some of the variance (Baayen, 2012). Nakagawa and colleagues (2017) have introduced a R-squared that represent the proportion of variance in generalized linear mixed-effect models. This method produces a marginal R-squared, presenting the variance explained by the fixed effects, and a conditional R-squared, which in addition presents the variance explained by the random effects. Overall, without the random effects accounted, the model accounted for almost 10% of the variance. After adding the random effects, the model explained 22.2% of the variance, which shows that the addition of random effects is an important addition for the explainability of the model.

Discussion

The goal of this research was to find relations between the impact of talk shows from the NPO and viewer engagement on social media. Natural language processing tasks based on machine learning algorithms were used to find tweet characteristics. These characteristics were added to a linear mixed model to find their relation to the impact scores of the three different talk shows. In this section, the results will be interpreted, implications in the research domain and ethics will be explained and finally, some limitations and suggestions for future research will be discussed.

Interpretation results

The results showed that the number of tweets per episode had a significant negative relation to the impact scores, which means that a higher number of tweets per episode, decreases the impact score. However, the change in the impact score is very small. If an episode scores one standard deviation above the average amount of tweets, the impact score decreases with -.434. This difference is close to zero and therefore does not have a large effect on the outcome.

The percentage of negative tweets is significantly negatively associated with the impact scores, which means that when there is a higher percentage of negative tweets compared to positive tweets about an episode, the impact score will decrease with -.068 per percentage. A higher percentage of positive tweets, therefore, has a positive effect on the impact score. This shows that more negative statements of the audience on social media are associated to a lower impact of talk show episodes.

There were a lot of tweets about other cases than the content and persons of the talk shows. When there is a higher percentage of tweets about the content of the talk show, the impact score increases significantly with 0.101 per percentage. This does have a large effect on the impact scores, considering that the average percentage of tweets about content lies around fifty percent. No significant result was found for the percentage of tweets about persons. Therefore, more valuable social media engagement, such as posts about the content of the television programmes, indicates that a talk show episode has a higher impact.

Finally, random intercepts were added for the different talk shows. There was more variance within talk show titles than between the different titles. However, the talk show titles did explain some of the differences in the impact scores. Altogether, the model had a conditional R squared of .222, which means that the model explained 22.2% of the variance. This result shows that the differences between television programme titles should be taken into account to get a more accurate prediction of impact scores.

Implications

The current research topic about the relation between viewer engagement and the impact of talk shows adds new information compared to previous research. The relation between tweets and viewing patterns has already been explored by Van Es and colleagues (2015). They researched whether viewing patterns could be explained by Twitter use and suggested that television programmes that spark polarized debates lead to more viewer engagement on social media. Nevertheless, the number of tweets was found to be distinct from the viewing patterns of television programmes.

The current research did search for a further connection between the experience of television viewers and viewer engagement on social media. However, due to the specific focus on PSM, the experience of viewers about public values, expressed in the NPO content, was included as the variable of interest. Different characteristics of tweets posted by viewers were used as a measure of online audience engagement. To find out whether this online user engagement could serve as a metric to evaluate whether the content fits to the goals of the NPO, this online engagement is related to the impact scores of talk shows which is an important measure from the NPO. As the results showed, the number of posts, sentiments, and the content of tweets together with the programme titles were related to the impact of television programmes. Consequently, this research adds knowledge to the domain of public service media and specifically to the relation online audience engagement on social media. However, more information about the relation is required. Many other social media engagement factors can potentially relate to the perceived impact of television programmes, for which further research is needed.

Ethical implications and consideration

When taking a look at the ethical implications of this research, several issues came up. First of all, participants generally give informed consent about their participation and should be able to stop participating at any moment, which was not possible during this research. Problems could occur when users deleted their posts. Because the tweets are already gathered, these tweets will still be included in the dataset. However, because the data is extracted from public Twitter profiles, the use of data by third parties and researchers is included in the terms and conditions from Twitter, which should be accepted by Twitter users. Users can accept these conditions without their full knowledge, due to a lack of comprehension or because they did not read it completely. This forms an ethical problem in terms of informed consent and could decrease feelings of autonomy and privacy of users. Nevertheless, the amount of information about Twitter users available in this research is limited. Only the @-mentions and retweet names were included in the non-cleaned corpora.

Besides the fact that Twitter users could feel a lack of autonomy when it comes to the usage of their tweets in academic research, no much harm is done to the participants. No tweets can be traced back to Twitter users and only the aggregated sum of tweets was used in the data analysis, which decreased the vulnerability of participants. Only the manual classification of the tweets can be seen as a potential risk of the vulnerability of users.

However, information about these tweets was kept private and was erased after the classification.

Finally, this research aims to get more information about the relation between PSM television programmes and online user engagement. The NPO wants to positively contribute to society and connect citizens. Therefore, there were no data used against participants. The only reason to use the data of viewers has been to reflect on the impact of television programmes and the associated social media engagement.

Limitations and suggestions for future research

Besides that this research has added more knowledge to the existing body of literature available in the public service media domain. There were some limitations found in the current research. First of all, the data was gathered via Twitter. Almost 40,000 different users tweeted about the three talk shows of interest. This active group of Twitter users that are expressing their opinions and participating in online discussions about the talk show is a specific group within society. It is difficult to generalize to the entire population when only the active social media users are included in the sample. This could bring potential external validation biases into this research and threaten the generalizability. However, this research does focus on online user engagement, which includes the risk of only measuring the experience of the online participating audience.

On the other hand, tweets only contain 280 characters, which gives limited space to users to express how they experience the talk show episodes. This might lead to more concrete posts, but also less specific posts. Because the tweets were classified based on their content, less relevant tweets were classified separately as *other*. In this way, only the tweets about the content of the talk shows were measured as predictors in the regression model.

Another limitation can be found in the limited amount of data that was available. Only impact scores rated by more than 30 participants were included, which makes the outcomes more reliable. However, all the episodes that were rated by less than 30 respondents were excluded from the data. In addition, tweets from longer than one year ago were unavailable and could not be included in the data. Unfortunately, this led to a reduced amount of information on which the linear mixed model was built. In the end, there was still information available about 343 talk show episodes, which is still a sample size about which conclusions can be drawn. For future research, it would be a great addition to measure viewers' engagement on other social media websites as well. Another improvement can be made if the people who rated the programs, were also measured on their audience engagement. In this way, the variance between persons would have been declined, which leads to more reliable conclusions.

Other limitations were introduced by the use of the variants of BERT classification and sentiment models. The models were not trained on a Twitter corpus, and even though the accuracy of the pre-trained model was good, there were still errors introduced into the dataset. In addition, the sentiments were labeled as either positive or negative, while there were a lot of tweets that could be assigned to neutral as well. The classified sentiments were checked and accepted as a valid method. In future research, better models could be created or the current models could be fine-tuned on other text corpora to increase their performances. In addition, more tweets should be pre-classified, to improve the accuracy of the robBERT classification model.

Besides, the linear mixed model does not prove a causal relation between the impact scores and social media engagement. Relations between the variables were discovered, but no specific information is available about the direction of this relation. While the impact scores were measured and based on feelings and thoughts during and after the programme episodes, the viewer engagement on Twitter was posted at the same time. An experimental design should be used to draw better conclusions about the causal relation between the variables.

Finally, this research created a basis for future research on the relation between viewers' engagement and the impact of television programmes. Future research should focus on the further evaluation of audience engagement as a metric, and find what other factors contribute to the impact scores of television programs, both online or in the characteristics of television programs themselves. In such a way, audience engagement could be used as an evaluation of the content of PSM content.

Conclusion

The current research is focused on the impact of Public Broadcaster talk shows from the NPO and the associated viewer engagement online. This research tries to discover whether online audience engagement online can evaluate the impact of talk show episode on humans. A sentiment analysis and text classification tasks extracted meaning from tweets, posted by viewers during three talk shows. This information was added to a linear mixed model in which the impact score was predicted. Results showed that there were some tweet characteristics associated with the impact scores of the talk shows. The number of tweets and the percentage of negative tweets led to a lower impact score, while the percentage of tweets about content led to a higher impact score. The talk show titles did explain some of the variance in the impact scores. More knowledge about potential predictors is needed to find out more about the relation between audience engagement and the content and impact of PSM television programmes. Therefore, future research is recommended.

References

Baayen, H. (2012). Mixed-effect models. Ln A. C. Cohn, C. Fougeron & M. K. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (2nd ed., pp. 668-677). Oxford University Press.

Bakker, P., & Scholten, O. (2003). *Communicatiekaart van Nederland: overzicht van media en communicatie.* Alphen aan den Rijn: Kluwer.

Bardoel, J., & Brants, K. (2003). *Public broadcasters and social responsibility in the Netherlands*. Nordicom Provides Information about Media and Communication Research, 167.

Bennett, J. (2013) The challenge of public service broadcasting in a web 2.0 era: broadcast production culture's view of interactive audiences. In Gulyás, A. & Hammer, F. (Eds.) *Public Service Media in the Digital Age: International Perspectives* (pp. 3-20). Cambridge University Press.

Benson, R., Powers, M., & Neff, T. (2017). Public media autonomy and accountability: best and worst policy practices in 12 leading democracies. *International Journa of Communication*, *11*, 1-22.

Bober, M. (2014). Twitter and TV events: and exploration of how to use social media for student-led research. *Journal of Information Management*, *66*(3), 297-312. https://doi.org/10.1108/AJIM-09-2013-0097

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, 4171-4186

Donders, K. (2019). Public service media beyond the digital hype: distribution strategies in a platform era. *Media, Culture & Society*, *41*(7). 1011-1028. https://doi.org/ 10.1177/0163443719857616

Doughty, M., Rowland, D., & Lawson, S. (2011). Co-viewing live TV with digital backchannel streams. *Proceedings of the 11th European Conference on Interactive TV and Video EuroITV '11, Lisbon, ortugal, ACM, New York, NY*, 141-144.

Doyle, G. (2010). From television to multi-platform: less from more or more for less? *Convergence*, *16*(4), 431-449. https://doi/org/10.1177/1354856510375145

Eke, C. I., Norman, A. A., & Shuib, L. (2021). Context-based feature techniques for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access*. *9*, 48501-48518. https://doi.org/10.1109/ACCESS.2021.3068323

Engelbert, J., & Awad, I. (2014). Securtizing cultural diversity: Dutch public broadcasting in post-multicultural and de-pillarized times. *Global Media and Communication*. https://doi.org/10.1177/1742766514552352

European Broadcasting Union. (2014). Public Service values, editorial principles and guidelines. https://www.ebu.ch/publications/strategic/login_only/guide/public-service-values-editorial-principles

García-Martínez, A. N., Nguyen, A. (2012). *When public service is the name of the game: the evolution of British and Spanish Public Service Broadcasters in online journalism*. In: Burns, M., Burgger, N. (eds.) Histories of Public Service Broadcasters on the Web. London: Peter Lang, pp. 1-16.

Gillan, J. (2011). *Television and New Media*, Routledge.

Highfield, T., Harrington, S., & Bruns, A. (2013). Twitter as a technology for audiencing and fandom. *Information, Communication & Society*, *16*(3), 315-339, https://doi.org/10.1080/1369118X.2012.756053

Huang, C. N., & Zhao, H. (2007). Chinese Word Segmentation: A Decade Review. *Journal of Chinese Information Processing*, *21*(3), 8-19. http://jcip.cipsc.org.cn/EN/Y2007/V21/I3/8

Kjus, Y. (2009). Impact of prestige programs on production practices: the case of cross-media and audience participation in public service organizations. *Journal of Media Practice*, *10*(2-3), 167-184, https://doi.org/10.1386/jmpr.10.2-3.167_1

Koerner, T. K., & Zhang, Y. (2017). Application of Linear Mixed Effects Models in Human Neuroscience Research: A Comparison with Pearson Correlation in Two Auditory Electrophysiology Studies. *Brain Science, 7*(3), 26. https://doi.org/10.3390/brainsci7030026

Luke, S. G. (2016). Evaluating significance in linear mixed-effects model in R. *Behavior Research Methods*, *49*, 1494-1502. https://doi.org/10.3758/s13428-016-0809-y

May, A. L. (2010). 'Who Tube?' How YouTube's News and Politics Space is going Mainstream. *International Journal of Press/Politics*. *15*, pp. 499-511. https://doi.org/10.1177/1940161210382861

Mediawet. (2008). https://wetten.overheid.nl/BWBR0025028/2018-05-30

Meijer, I. C. (2005). Impact or content? Ratings vs Quality in Public Broadcasting. *European Journal of Communication*, *20*(1), 27-53. https://doi.org/10.1177/0267323105049632

Miles, J., & Shevlin, M. (2001). *Applying regression & correlation: a guide for students and researchers* (First ed.). SAGE Publications Ltd.

Murdock, G. (2018). *Reclaiming Digital Space. From commercial enclosure to the broadcast commons*. In Lowe, G. F., Van den Bulck, H., Donders, K. (Eds.), Public service media in the networked society (pp.4-58). Nordicom.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*, 1-11. https://doi/10.1098/rsif.2017.0213

NPO-Organisatie. (2020). Concessiebeleidsplan 2022-2026. https://over.npo.nl/concessiebeleidsplan

NPO Publieksonderzoek 2021 (2021). Impact beleving instrument. https://over.npo.nl/download/nl/36****

NPO. Het Publieke bestel. https://over.npo.nl/organisatie/bestuur-en-organisatie/het-publieke-bestel#content

Ortiz Suarez, P. J., Sagot, B., & Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why Copyright © 2004. researchers should ALWAYS check for them). Practical Assessment, Research, All rights reserved and Evaluation, 9(6) Online at http://pareonline.net/getvn.asp?v=9&n=6

Oudeman, M., Albrecht, Y., van Bockxmeer, H., Dibbits, T., Helberger, N., Kasem, I., Vanseveren, W., & Vermeir, L. (2019). Samen voor het publiek belang: Evaluatierapport NPO 2013 t/m 2017. https://hdl.handle.net/11245.1/ccf1adc1-032b-4c45-ac21-28409b8c5684

Page, R. (2017). *Narration*. C Hoffmann & W Bublitz (eds.), *Handbook of Pragmatics 11: Pragmatics of Social Media.* De Gruyter, Berlin, pp. 523-544. https://doi.org/10.1515/9783110431070

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*, *10*, 1320-1326.

Picard, R. G. (2002). Assessing Audience Performance of Public Service Broadcasters. *European Journal of Communication*, *17*(2), 227–235. https://doi.org/10.1177/0267323102017002696

Proulx, M., & Shepatin, S. (2012). Social TV: How Marketers can reach and engage audiences by social media, and mobile. Hoboken, NJ: John Wiley & Sons.

Schwarz, J. A. (2016). Public Service Broadcasting and Data-Driven Personalization: A View from Sweden. *Television & New Media*, *17*(2), 124-141. https://doi.org/10.1177/1527476415616193

Sjøvaag, H., Stavelin, E., & Moe, H. (2016). Continuity and change in public service news online: A longitudinal analysis of the Norwegian Broadcasting Corporation. *Journalism Studies*, *17*(8), 952-970. https://doi.org/10.1080/1461670X.2015.1022204

Sjøvaag, H., Pedersen, T. A., & Owren, T. (2018). Is public service broadcasting a threat to commercial media? *Media, Culture & Society*, *41*(6). https://doi.org/10.1177/0163443718818354

Slob, A. (2019, June 14). *Visie toekomst publiek omroepbestel: waarde voor het publiek* [Letter of government]. Retrieved from https://www.tweedekamer.nl/kamerstukken/detail?id=2019Z12224&did=2019D25185

Syvertsen, S. (2003). Challenges to Public Television in the Era of Convergence and Commercialization. *Television & New Media*, *4*(2), 155-175. https://doi.org/10.1177/1527476402250683

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406-418. https://doi.org/10.1002/asi.21462

Twitter Developer. (2020). Twitter Development Agreement and Policy. https://developer.twitter.com/en/developer-terms/agreement-and-policy

Van Dijck, J., & Poell, T. (2014). Making Public Television Social? Public Service Broadcasting and the Challenge sof Social Media. *Television & New Media, 16*(2), 148-164. https://doi.org/https://doi.org/10.1177/1527476414527136

Van Es, K. F. (2017). An impending crisis of Imagination: Data-driven personalization in public service broadcasters. *43*, 1-18.

Van Es, K. F. (2016). *The Future of Live*. Wiley

Van Es, K. F., Van Geenen, D., & Boeschoten, T. (2016). Re-imagining Television Audience Research: Tracing Viewing Patterns on Twitter. *M/C Journal*, *18*(6). https://doi.org/10.5204/mcj.1032

Van Es, K. F., & Poell, T. (2020). Platform Imaginaries and Dutch Public Service Media. *Social Media & Society*. 1-10. https://doi.org/10.1177/2056305120933289

Van der Burgh, B., & Verberne, S. (2019). The merits of universal language model fine-tuning for small datasets–a case with Dutch book reviews. arXiv preprint 1910.00896.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*

Vyas, V., & Uma, V. (2019). Approaches to sentiment analysis on product reviews. *Sentiment analysis and knowledge discovery in contemporary bursiness*. Hershey, PA, USA: IGI Global, 15-30.

Wang, Y., Zhou, Z., Jin, S., Liu, D., & Lu, M. (2017). Comparisons and Selections of Features and Classifiers for Short Text Classification. *IOP Conference Series: Materials Science and Engineering*, *261*. https://doi.org/10.1088/1757-899X/261/1/012018

Williams, G. (2018). *R for Psych*. https://glennwilliams.me/r4psych/

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140