Predicting Online Participation in Public Broadcasting Using Machine Learning

MSc. Applied Data Science

Prepared for Dr. Mirko Schaefer, Dr. Max Kemman, Joris Veerbeek

Wouter Regter

0961817

02-07-2021

ABSTRACT

As it remains unclear how the success of public broadcasting on social media should be measured, this thesis argues that online participation could serve as an alternative, social media metric that aligns with public broadcasting's traditional aims and draws support from more general public research and audience research. Consequently, public broadcasting benefits from understanding what online participation can be expected of their new media content. For this reason, and to help characterize online participation as an alternative metric of online participation, this thesis aims to investigate to what extent a predictive model of online participation can be built, and which predictors are important in the process. In addition, this thesis aims to investigate whether topic modeling can be applied on the subtitles of public broadcasting shows to generate useful features for the prediction of online participation. Using data from the Dutch public broadcasting service, 22 potentially predictive features of online participation were collected and created, missing values and outliers were dealt with, and 7 models were individually tuned and compared with the aim of achieving the best prediction performance. The results suggest that although most values of online participation can be predicted with decent accuracy, the model performs poorly on large values of online participation. Furthermore, the results indicate that the inclusion of topics as features did not lead to significant improvements in prediction performance but do generate some useful insights. Scholars and public broadcasting organizations may use the results of this thesis to enhance their understanding of online participation as an alternative metric of broadcasting success.

INTRODUCTION

Social Media and Public Broadcasting

Over the last decade, social media platforms have transformed social interactions, making information easily shareable and accessible (Chen et al., 2011). Social media platforms are driven by Web 2.0 technology, a technical infrastructure that facilitates user-centered design, collaboration, and information sharing (Power & Phillips-Wren, 2011). These affordances of Web 2.0 technology have attracted many organizations to participate on social media platforms because it enhances their interactions with the groups of individuals they are trying to reach, and the interactions between those individuals (Ksiazek et al., 2016), leading to a variety of benefits in a diverse set of contexts. In the public sector, scholars found that participating on social media can help extend government services (Bertot et al., 2012), increase civic participation (Rice et al., 2013; Zhang et al., 2010) and enquire citizen's opinions and behaviors (Chun & Reyes, 2012). In the private sector, tourism management scholars found that social media platforms were an effective means for promoting products and services (Harrigan et al., 2017). Furthermore, marketing management scholars found that participating on social media can help to spread electronic word-of-mouth (Hudson et al., 2016), and manage customer relations (Coulter & Roggeveen, 2012). As the public and private sectors' aims are inherently different, they use social media in different ways to create different benefits. However, particularly in public service broadcasting, critics have noted that the encounter between public aims and social media has, in some cases, led to the compromise of public values (Bardoel & Lowe, 2007; Van Dijck & Poell, 2015).

Due to the rise of social media platforms such as YouTube, Twitter, and Facebook, both public and commercial broadcasting are increasingly incorporating social media features in their more traditional mass media approach (Van Dijck & Poell, 2013). Consequently, likes, views and followers are becoming gradually more important indicators of broadcasting success and are internalized by both commercial and public broadcasting (Van Dijck & Poell, 2015). A result of the commercial nature of current-day social media platforms is that the contents on social media platforms carry commercial value. For example, a YouTube video generates advertising income based on views and likes. Moreover, the content that is likely to yield many likes or views is often not in line with public broadcasting's traditional values (Enli, 2008). Hence, the extent to which public broadcasting should engage on social media platforms is debated. Furthermore, in the past, public broadcasting has been criticized for compromising public value in the pursuit to stay competitive when commercial broadcasting, and later, social media platforms started competing in the space (Bardoel & Lowe, 2007). Companies like

Youtube, Twitter and Facebook have a clear advantage over public broadcasting as they can make enticing offers to independent creators that include monetizing their content through targeted advertising and distributing it globally (May, 2010). Hence, public broadcasting struggles to pursue its goals as a public organization while remaining relevant in today's media landscape. As a result, scholars have suggested that, with the developments in technology and the resulting changes in the media landscape, public service broadcasting should be redefined (Enli, 2008; Van Dijck & Poell, 2015; Bardoel & Lowe, 2007).

One of the expectations of public broadcasting is that, rather than seeing viewers as consumers that consume content, participation is prioritized over consumption and viewers are seen as participative citizens (Van Dijck & Poell, 2015). Related to this view, public broadcasting's success is expressed "not in the time people spend watching the programs, but the time spent elsewhere, exercising what they have learnt" (Syvertsen, 2004, p. 367). In addition, since public broadcasting's emergence, its widely accepted core principles have been to "entertain, inform and educate" (Holtz-Bacha & Norris, 2000). In the pursuit to redefine public broadcasting while remaining true to its core mission and principles, several scholars propose the change of 'public service broadcasting' to 'public service media' (Murdock, 2005; Moe, 2008; Coleman, 2004; Bardoel & Lowe, 2007). These scholars argue that public broadcasting should go beyond television and embrace the full possibilities of the internet, including social media, because the internet can be utilized to engage citizens, while the core values of public service broadcasting are retained. But with this line of argument, it remains unclear how public broadcasting success can be measured, especially considering the apparent misalignment between public broadcasting's values and social media's collection of metrics such as views and likes.

When considering public broadcasting's mission to engage citizens to participate, we may draw from the body of literature that researches 'citizen engagement', a concept that is not specifically used in the context of public broadcasting but in a broader public sense. Although citizen engagement is a wide concept with no clear consensus on its definition, the term 'engagement' with respect to the public sector has been defined as the participation of citizens in social issues within the community (Gil de Zúñiga et al., 2012), but more fundamentally, it represents the interaction between government and citizens (Ekman & Amn, 2012). Citizen engagement has been considered an important aspect of creating public value (Benington, 2011). Some scholars noted that social media can be considered a medium for encouraging citizen participation through collaboration, debate, and coordination of public and social activities (Criado et al., 2013; Warren et al., 2014). Allowing citizens to contribute

content to social media channels of the government or enabling them to share their thoughts online has been found an effective way to encourage online participation and thereby improve citizen engagement (Mergel, 2013). It is difficult to translate citizen engagement directly to the context of public broadcasting because there are some important differences between the government in general and public broadcasting, specifically in their interaction with citizens. While the aim of the interaction between the government in general and citizens is to address social/political issues and inspire debate on these issues (Gil de Zúñiga et al., 2012), the purpose of the interaction between public broadcasting and citizens is to entertain, inform and educate (Holtz-Bacha & Norris, 2000). Although these two goals overlap, public broadcasting's aim is more focused on engaging citizens in a broader sense, where participation of citizens is not strictly related to social/political issues. Hence, the goal of citizen engagement in public broadcasting is to engage citizens to participate with public broadcasting content that is meant to entertain, inform, and educate. Consequently, as online participation is considered to be, although in a broader public context, an important driver of engaging citizens (Mergel, 2013), online participation should be considered an important indicator of public broadcasting success.

Social Media Metrics of Broadcasting Success

Online participation is a concept measured through social media metrics, mostly through comments or responses on posts of online pages (e.g. Wang & Bryer, 2013; Sukumaran et al., 2011). Hence, online participation can be considered a type of social media metric. Furthermore, as was argued before, in the context of public broadcasting, online participation can be considered an indicator, or alternative metric, of public broadcasting success, because of its relation to citizen engagement. Van Es et al. (2016) have critically examined the usefulness and validity of broadcasting's social media metrics in the context of audience engagement. Van Es et al. found that as social media "invite new forms of inquiry", social media metrics can be used to create useful insights, but also have some limitations that should be considered. Furthermore, Van Es et al. argue that research on broadcasting metrics should "follow the medium", by preferring metrics related to the medium over traditional metrics, as this would allow researchers to utilize the new forms of inquiry of social media. Therefore, by examining online participation, which is essentially a type of social media metric, in the context of public broadcasting, new forms of inquiry of social media can be utilized to create useful insights for public broadcasting.

Hence, as increasing the understanding of online participation in this setting can create useful insights for public broadcasting, this thesis aims to analyze online participation in this

setting, thereby contributing to the debate on public values and social media in public broadcasting (Van Dijck & Poell, 2015), by enhancing the understanding of online participation, an alternative metric of broadcasting success and a driver of citizen engagement (Mergel, 2013), one of public broadcasting's main aims.

When examining social media metrics such as online participation in public broadcasting, there are several limitations to using social media metrics that are relevant, and these should be carefully considered. In the context of media audiences, Baym (2013) proposed two important limitations that are important to consider when examining online participation as an alternative metric of public broadcasting success. Baym proposed that social media metrics are skewed due to algorithmic ranking of posts, meaning posts with higher engagement are more likely to receive more engagement because they are ranked higher by the algorithm and are shown more often to users. For example, if a post placed on the social media channel of a television show initially receives some comments, it becomes more likely that others will also see the post and start to comment, thereby skewing online participation through the dynamics of the social media platform itself. Furthermore, Baym proposed that social media metrics "are non–representative samples of audiences and their engagement". For example, the sum of comments on a single or even multiple social media channels of television shows is not representative of all viewers as a whole. This means that, when measuring online participation in public broadcasting, there is an inference being made on a non-representative sample of the viewers of public broadcasting, meaning the results should be interpreted with caution.

Furthermore, with respect to the limitations of social media metrics, Gerlitz (2017) put forward a data point critique and argued that social media counts composed of very dissimilar elements are often too straightforwardly interpreted as a count of homogeneous elements. As a results, Gerlitz suggested treating social media indicators not as first-order metrics but as "already assembled second-order metrics, which are composed of heterogeneous interpretations, practices and actors". Therefore, Gerlitz argued that it is important to understand what is being counted before composing a metric. Hence, data points such as the number of comments should not be straightforwardly interpreted as a composition of similar comments, but as a composition of comments that differ in interpretation, practices, and actors. Furthermore, it is important to understand how these comments differ before considering them as a metric.

Online Participation and Public Broadcasting

So far this thesis has focused on discussing the encounter between public broadcasting and social media, and the relevance of online participation as an alternative metric of broadcasting success. As this is a problem associated with media science and related humanities research, most of the literature that was introduced originated from this domain. However, online participation is a multidisciplinary concept, and the related literature takes up a more social science perspective. The body of literature that examines online participation, or active participation, as it is sometimes referred to, is extensive (e.g. Malinen, 2015; Livingstone et al., 2005), but does not cover online participation in the context of public broadcasting. We cannot expect the research on online participation to generalize well to the context of public broadcasting because the research mostly covers online participation in response to online posts, while in the context of public broadcasting, the online participation of viewers is in response to online posts related to a program the viewer watched. Consequently, program features are neglected. Program features are important to consider in the context of public broadcasting because we expect that different program features lead to different extents of online participation. For example, we can expect there to be a difference in the extent of online participation between programs that cover controversial topics like climate change and programs that cover less controversial topics but are equally popular. Thus, as the literature does not cover online participation in the context of public broadcasting, and it cannot be expected that the current online participation literature can generalize well to this context, online participation in public broadcasting is not well understood. Therefore, by aiming to increase the understanding of online participation in this context, this thesis contributes to the literature regarding online participation (e.g. Malinen, 2015; Livingstone et al., 2005) by extending online participation to the context of public broadcasting, thereby considering program features in the analysis of online participation.

As discussed so far, there is a need to understand online participation in the context of public broadcasting, which means program features should be considered. A mathematical tool that can help capture the complex characteristics of the content of television programs is topic modeling. A topic model is a statistical model which has the goal of finding short descriptions (i.e. topics) of large amounts of text efficiently while preserving the essential statistical information that is useful for tasks like classification and summarization (Blei et al., 2003). Topic modeling has several applications with regards to social media, many of them using tweets on Twitter to, for example, examine the topics of user opinions (e.g. Ramamonjisoa, 2014; Hu & Ester, 2013) or the topics of news streams (e.g. Zhao et al., 2011; Gao et al., 2018).

The subtitles of television programs are large amounts of text that could, using a topic model, be efficiently processed to generate useful statistical information related to the topics. Furthermore, we can expect that different program topics lead to different extents of online participation. This is also illustrated by the example that was mentioned earlier; controversial topics can be expected to generate more online participation than less controversial topics. Therefore, we can expect a topic to be a useful predictor of online participation in public broadcasting. By increasing the understanding of how topic modeling can be used in the setting of public broadcasting, scholars in this field will benefit from this by increasing the range of applications in which topic modeling can be effectively applied. Furthermore, it is relevant to understand this because public broadcasting may use topic modeling to enhance their prediction of online participation, or they may be able to identify topics that are likely to generate high online participation. Thus, this thesis aims to contribute to the literature on topic modeling and its applications by presenting a new application of topic modeling, specifically by applying it to the subtitles of television programs to predict online participation.

As discussed previously, online participation can be considered an alternative metric for public broadcasting success. Hence, public broadcasting benefits from understanding what online participation can be expected of their new media content, so that they can adjust their media content in response to this prediction or make other adjustments such as changing their social media campaign. Moreover, public broadcasting has large data resources available, that can be used to train predictive machine learning models and potentially generate promising predictions of online participation. Furthermore, understanding what predictors are important in the prediction of online participation helps them identify which types of programs generate high online participation. Therefore, this thesis aims to investigate to what extent a predictive model of online participation can be built, and which predictors are important in the process. The resulting findings help characterize online participation as an alternative metric of online participation, and the application of online participation in the context of public broadcasting. To do this, this thesis uses the Dutch public broadcasting service, or the Nederlandse Publieke Omroep (NPO), as an empirical example. In this thesis, the data of the NPO is used to create a diverse set of features, and train several machine learning models, aimed at making the best predictions of online participation. In addition, as topic modelling is likely a promising method to extract predictive features from the subtitles of television programs, this thesis aims to incorporate the features generated from a topic model to see if they are useful predictors in modelling online participation. Hence, the main question this thesis aims to answer is: ‘What useful insights can be extracted from a predictive model of online participation as an alternative

metric of broadcasting success?' Furthermore, to answer this question, the sub-questions that this thesis aims to answer are:

(1) To what extent can online participation be predicted in the context of public broadcasting?
(2) What are the most important features when predicting online participation in the context of public broadcasting?
(3) Can useful features of online participation be extracted using topic modeling on the subtitles of programs in the context of public broadcasting?
(4) Can topics be identified that are likely to generate high online participation on social media in the context of public broadcasting?

This thesis makes several theoretical contributions. Firstly, the results of this thesis add to the debate on public values and social media in public broadcasting (Van Dijck & Poell, 2015), by enhancing the understanding of online participation, a driver of citizen engagement (Mergel, 2013), one of public broadcasting's main goals. Secondly, the results of this thesis contribute to the body of literature regarding online participation (e.g. Malinen, 2015; Livingstone et al., 2005) by extending online participation to the context of public broadcasting, thereby considering program features in the analysis of online participation. Thirdly, the results of this thesis add to the literature on topic modeling and its applications (e.g. Ramamonjisoa, 2014; Gao et al., 2018) by presenting a new application of topic modeling, specifically by applying it to the subtitles of television programs to predict online participation. Furthermore, the results of this thesis are practically relevant. This thesis attempts to predict online participation based on a diverse set of features based on the large data resources that public broadcasting has. Public broadcasters may adopt a similar model to predict the expected online participation of their programs or may consider using the features that showed to be important for the performance of the model. This allows them to better understand online participation and potentially make decisions to generate more online participation. This is especially relevant because online participation can be considered an alternative metric of broadcasting success.

DATA

Data Collection

Due to arrangements with the NPO, I was given access to their Google Cloud BigQuery databases from which the relevant data could be collected using SQL. The data that was used for answering the research questions consists of episode related data and daily social media data. The episodes to include in the analysis were selected based on several criteria. Firstly, the program that released the episode must have a social media channel and have social media data available. If this were not the case, online participation could not be calculated. Secondly, only the first releases of episodes were included, as repetitions generally do not generate much social media engagement and the calculation of online participation around the release date of the episode could interfere with the calculation of online participation around the release date of new releases. For example, if there is a repetition of a show on the next day after the new release of the episode, much of the online participation may be attributed to the repetition of the episode while these comments are most likely in response to the newly released episode. In total, 3204 episodes were collected.

The social media platform that was chosen to measure online participation is Facebook. On Facebook, discussion about the episodes is generated that is reflective of the concept of online participation in this context. Motivated through Gerlitz's (2017) data point critique, it is important to understand what is being counted before composing a metric, because "social media metrics are composed of heterogeneous interpretations". Although only the number of comments on posts are available in the database and not the comments themselves or the replies on those comments, the following examples were retrieved from the Facebook page itself to illustrate the commenting behavior on Facebook. On the Facebook channel of 'Wie is de Mol', an amusement program in which contestants try to uncover who among them is secretly sabotaging the assignments they must fulfil to receive money, users on Facebook discuss who acted suspiciously during the episode in the comments on Facebook. For example, on a post that received 452 comments and was placed on the 27th of February 2021 right after an episode, one user replied: "Sportswoman Rocky chose to not run 1 kilometer to the finish while that would yield money and at the beginning of the assignment, she said that she would like to earn money". Rocky is one of the contestants in the show and the user commented that she was acting suspiciously because in the user's opinion, she was not doing everything she could to earn money, which is the goal of the contestants. The comment received 14 responses. Furthermore, in another comment on the same post, another user posted: "Too bad that there weren't three finalists in this episode, this made it way less exciting, weird decision of the

program makers". In this comment, a user expressed his/her opinion about the content of the episode. These comments can be considered valuable for public broadcasting because users are engaging with the media content of public broadcasting, through participating online in the discussion of the content of the show, and through sharing their opinions about the show itself. In addition, these forms of participation are valuable because they indicate the formation of communities on the NPO's social media channels.

Other comments can be considered less valuable, which seem to be posted by Facebook users who did not see the episode. For example, on a post placed on the 28th of February 2021, a day after the episode, one user replied that he could not see the episode because he did not have time to watch it yet. These types of comments are less valuable to public broadcasting because this user does not engage with the media content of the NPO, the user just felt the need to reply to the post for other reasons. However, there are usually only a few of these comments on posts close to an episode and most of the comments on posts close to the episode are about the content of the episode. However, as the time distance between post and episode increases, the comments are less about the content of a specific episode, and more about the content of the post itself, which is something that should be considered. For example, sometimes there are videos placed in a post on the social media channels of shows that present extra content unrelated to the episode, or there is a post placed with a certain question for the viewers, that may be unrelated to the content of an episode. However, this usually occurs more frequently when the time distance between post and episode increases. Hence, this can be accounted for by weighing the number of comments by the time distance.

Considering that these comments are counted and straightforwardly composed into a metric without identifying the individual differences between these comments imposes its limitations on measuring online participation as a metric of alternative broadcasting success (Gerlitz, 2017). As demonstrated in the last two paragraphs, there are inherent differences between comments that are not considered in the comment count. For example, some users comment about the episode while others do not, and some comments receive more replies than others. These are not the only relevant differences, some other examples include the time difference between the post and the comment, the intention of the comment and characteristics of the user such as the average online activity of the user. By composing a metric without considering these differences, important information is lost that could help make the count of comments more reflective of online participation, by, for example, only considering comments that are about the episode. At the least, it could help explain the characteristics of comments

that are taken as a sample. Anyhow, it is important to consider that this therefore limits the measurement of online participation in these ways.

Subtitle Selection

The subtitles of the episodes of the NPO shows are all generated through automatic dictating software and are manually corrected by the producer if necessary. Furthermore, the subtitles of live episodes, as opposed to pre-recorded episodes, are usually dictated by the software during the live episode. In terms of reliability, the pre-recorded shows show very little error, with usually only a few grammatical errors. However, live episodes that are dictated during the episode generally contain more error due to falsely dictated words or inaudible speakers. Overall, live episodes make up approximately 40% of the total episodes. Two sets of subtitles of the shows of the NPO were collected and all of them are in Dutch. Firstly, the set of subtitles that are of the episodes that have been discussed so far and adhere to the criteria discussed at the beginning of this section. These are the episodes for which the topics need to be found in order to answer the research questions. Secondly, a larger set of subtitles was collected for training the topic model. The episodes included in this larger set of subtitles do not necessarily have to adhere to the criteria discussed earlier. For example, the shows that these subtitles are of, do not need to have a social media channel. Hence, by not limiting the set of subtitles to these criteria, the set is much larger, containing 44071 subtitles instead of 3204. This set of subtitles is used to train the topic model, because with more data, it is likely that the model can generate more well-defined topics for the episodes that are important for answering the research questions.

Ethical Considerations

Most of the episode related data captures the characteristics of an episode and the reception of viewers. The data that captures the reception of viewers, such as the viewer count of a specific episode, does not contain any personal information of viewers because the data in the database is aggregated at episode level. Hence, the viewing numbers of specific individuals are not accessible in the database and the behavior of specific individuals is not being tracked. In addition, the social media data tracks the weekly and daily activity on the NPO's channels on social media platforms. This includes indicators like the number of likes, comments, and views on the social media pages of the NPO in general and on the posts placed on those pages. However, this information is not collected on an individual level. The social media activity is tracked through the aggregation at post or page level. Hence, for a post or page, the number of comments on certain day or week is collected, yet, the content of a specific comment, or which

user posted the comment is not collected. Thus, as both the episode related data and social media data do not track the individual behavior of viewers and no personally identifiable information is being collected, the risks of misuse of personal information of users are mitigated.

METHODS

Formulating the Prediction Problem

The code that was used for all operations described in this section can be found in this repository: https://github.com/wouterregter/thesis_ads. This thesis has two aims and for these aims, the methods are discussed separately. Firstly, the methods are discussed for the prediction of online participation without the inclusion of topics generated from topic modeling, and, subsequently, the methods for topic modeling and the analysis of the topics are discussed. The first aim of this thesis is to answer to what extent online participation can be predicted in the context of public broadcasting and what predictors are important. As the target for prediction is online participation, measured by the number of comments on Facebook posts, the target is numerical. Hence, in the context of supervised machine learning, the question introduces a regression problem rather than a classification problem. Furthermore, when referring to the context of public broadcasting in terms of this regression problem, features should be selected or created that are predictive of online participation. As the R^2 and *RMSE* of a regression model indicate how well a variable can be predicted based on a set of features, these performance measures can help determine to what extent online participation can be predicted. Furthermore, as the most important features contribute most to the prediction of the target, this can be used to evaluate which features are important. Thus, translating the research questions to data science questions, this thesis aims to answer: (1) *'What R^2 and RMSE can be achieved on a predictive regression model with online participation as a target?'* and, (2) '*Which features contribute the most to the performance of the model?'.* The methods to answer these questions are discussed next, starting with the variable selection and creation process, followed by data preparation and the modeling approach.

Variable Selection and Creation

Online Participation. Online participation, the value that is the target for prediction, was calculated through the comments on Facebook posts. First, the distances between the release date/time of the post and the release date/time of all first released episodes of a show were calculated and the closest episode was linked to each post. This was done to ensure that the comments are attributed to the right episode only, and not multiple times when there are multiple episodes in a short time. Subsequently, a set of weights were multiplied with the number of comments of the posts, depending on how close the post is to the closest episode. A weight of 1 was multiplied with the number of comments on posts that were 24 hours before or after the episode, a weight of 0.5 for 48 hours before or after the episode, a weight of 0.25 for 72 hours before or after the episode and a weight of 0 for more than 72 hours before or after

the episode. This was done to ensure that only the comments that are close to the episode are included, and therefore are likely to be about the episode. The weights were chosen based on domain knowledge of the employees of the NPO. Subsequently, all posts were aggregated to episode level by summing all weighted number of comments, yielding the total sum of weighted comments per episode. The calculation of online participation in this way assumes that the comments placed on posts occur only within a limited time after the placement of the post, which is most often the case.

Episode Characteristics. Characteristics of episodes are likely to influence online participation. For example, some channels of the NPO may generally have more viewers, hence, more viewers may have something to discuss about the episode online. Therefore, if an episode is released on a certain channel, there may be a difference in online participation because it was released on that channel. 12 features that characterize an episode were included, these were directly available from NPO's databases: (1) Title, (2) Channel, (3) SKO Class L1, (4) SKO Class L2 (5) CCC Domain Definition, (6) CCC Definition, (7) Zapp, (8) Broadcasters, (9) Minimum Age Classification, (10) Duration, (11) First Broadcaster, and (12) NOS Content. *Title* is categorical and represents the title of the show. *Channel* is categorical and represents the television channel that the episode was released on. The features *SKO Class L1, SKO Class L2, CCC Domain Definition, CCC Definition* are categorical and represent the content categories of the episode. For example, the features indicate whether an episode is fiction or non-fiction and whether it is about actualities or amusement. The feature *Zapp* is Boolean and represents whether the episode is part of NPO's television channel aimed at teenagers. The feature *Broadcasters* is categorical and represents which broadcasters released the episode. Unlike most other countries' broadcasting organizations, the NPO consist of multiple member-based broadcasting associations, each representing different political and religious views. The feature *Minimum Age Classification* represents the advised minimum age for viewing the episode. The feature *Duration* is numerical and represents the length of the episode in milliseconds. The feature *First Broadcaster* is categorical and represents which broadcaster released the episode first. The feature *NOS Content* is Boolean and represents whether the episode is part of NPO's main news channel.

Two other features were calculated that are related to episode characteristics: Start Season/Series (13) and Historical Average Views (14). *Start Season/Series* is categorical and indicates whether the episode is the start of a season, the start of a series, or no start of a season/series. *Historical Average Views* represents the historical average viewer count of programs released over a year ending two months before the release date of the episode. Hence,

the historical average view count is calculated by considering programs that are released between 14 and 2 months before the release date of the episode. The reason for this is that two months before the release of the episode, the NPO issues a release plan with the characteristics of the season/episode. Hence, at this point, the release date is certain and the information that is used as features should be available. Therefore, this is the ideal moment to predict the online participation of an episode for most shows, as this is the point at which the information about the episode becomes available while there are still two months to make potential adjustments to the episode. Therefore, a year ending two months before the release was taken because at two months before the release of the episode, the daily historical social media engagement over the last year ends at that point. A period of a year was averaged over as most shows release at least one episode once a year, but two seasons may be in the first and last quarter of the year.

Historical Social Media Engagement. Previous engagement on social media can be expected to influence online participation. For example, fans on Facebook receive updates in their news feed about new posts of the channel, which makes it more likely that they will see the post and comment on it. Hence, the more fans, the higher the expected online participation. Historical social media engagement was calculated by taking the daily average of several indicators over a year ending two months before the release date of the episode, for the same reasons discussed in the previous section. The NPO has many social media engagement indicators available but only the Facebook indicators were considered, as these should be most predictive. Furthermore, among the Facebook indicators there exists high multicollinearity, for example, the average number of likes highly correlates with the average number of comments of the social media channel. Therefore, the Facebook indicators that correlated most with the target were retained, and others were dropped until no Pearson correlation coefficients exceeded 0.80.

The features that were retained are (15) Fans, (16) Fans Change, (17) Own Posts, (18) Own Posts Comments, (19) Own Posts Engagement Rate, (20) Page Impressions Unique, (21) Page Impressions Paid Unique and (22) Page Views Total. All the features are numerical. The feature *Fans* represents the average daily number of page likes, *Fans Change* represents the average daily change in page likes, *Own Posts* represents the average number of daily posts released on the social media channel, *Own Posts Comments* represents the average daily number of comments on the posts of the social media channel, *Own Posts Engagement Rate* represents the average daily number of comments, shares, and likes on posts divided by the number of people who had any content from the page enter their screen, *Page Impressions Unique* represents the daily average number of people who had any content from the page enter

their screen through paid distribution such as an ad, *Page Views Total* represent the daily average number of times a page's profile has been viewed.

Data Preparation

Figure 1. Histogram of Online Participation

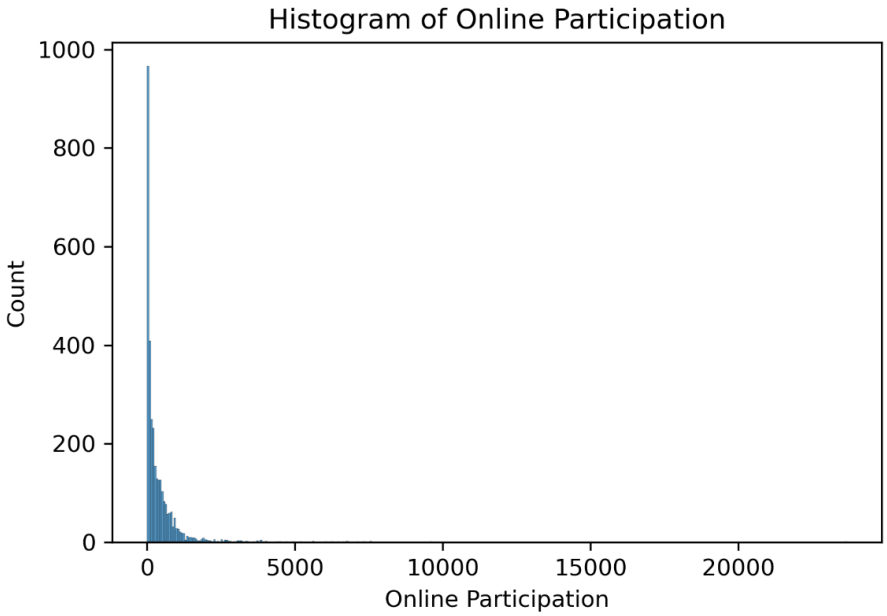
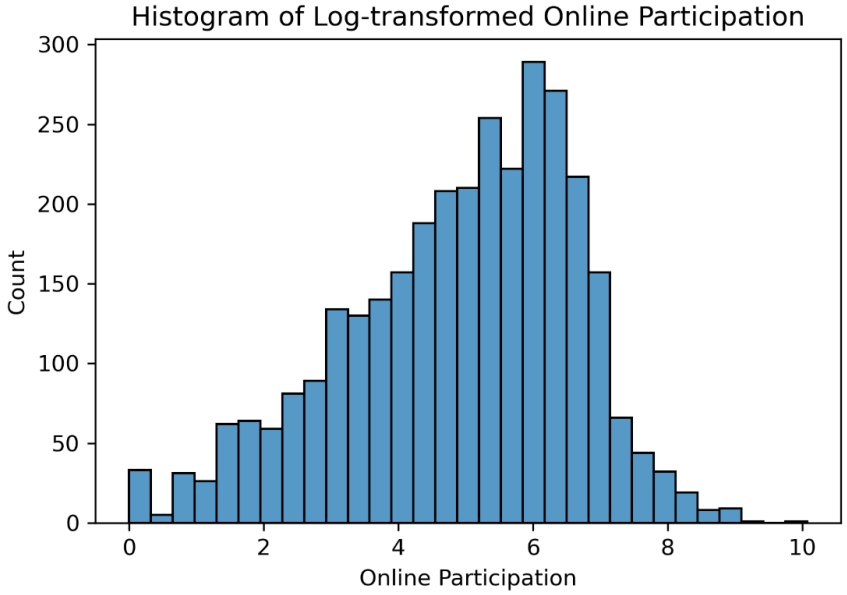


Figure 2. Histogram of Log-Transformed Online Participation



Outliers. As shown in Figure 1, most values of online participation are in the range 0-2500, but there are also a few that exceed 20000 that are not even visible in the Figure. This is

a problem because the few very high values likely distort the prediction models, because the models will focus on reducing the large errors for high values of online participation, leading to worse predictions of the lower values. Removing the outliers is not ideal because unseen examples with very large numbers of online participation will yield inaccurate predictions of online participation when excluding them in the model. Hence, a log transformation was applied to deal with the outliers and skewness of online participation. Specifically, $y' = \ln(y) + 1$, where y is online participation, was taken because there are some zeros in the data. The resulting distribution, shown in Figure 2, is much less skewed and contains no outliers.

Missing values. There were several features with missing values. The categorical feature *Minimum Age Classification* contained many missing values because a missing value indicated that there was no minimum age classification for that episode. Hence, these missing values were replaced with the category 'None'. The features *Fans* and *Fans Change* consisted of missing values because for the Facebook data, the data was collected through two APIs that in some ways overlap but in other ways do not. In this case, two other features represented the same features as *Fans* and *Fans Change* but were collected through the other API. Hence, these values could be directly imputed from those features. *Page Engagement Rate Unique*, *Page Views Total*, *Page Impressions Paid Unique* and *Own Posts Engagement Rate* also contained some missing values, caused by the difference in Facebook APIs, but could not be imputed by other data in the database. A logistic regression model was fit to predict whether the values for these variables were missing or not, based on the other features. With an accuracy of 91.3%, it was concluded that the missingness could be predicted based on the available data, indicating the data is Missing At Random (MAR). Therefore, the missing values for these features were imputed using scikit-learn's Iterative Imputer, which imputes missing values by modeling each feature with missing values as a function of other features, using a Bayesian Ridge model. After the removal of missing values and the creation of dummy variables from the categorical variables, the data consisted of 3204 records and 118 features.

Modeling Approach

Several regression models were selected, tuned, and compared using Python (version 3.8.8), the scikit-learn library (version 0.24.0) and the xgboost library (version 1.5.0). The selection of model includes (1) Linear Regression, (2) Ridge Regression, (3) LASSO Regression, (4) Elastic Net Regression, (5) Support Vector Regression, (6) Random Forest Regression and (7) Extreme Gradient Boosting Regression. The models were chosen on the basis that they are well-established in the field and have demonstrated their usefulness in regression problems (Bonaccorso, 2018). Some more complex models like neural networks

have been excluded from the selection because determining feature importances of these models did not prove to be computationally feasible with the available computational resources. Train and test sets were created using 25% of the data for the test set, and a random state was set for a fair comparison between results. It has to be noted that when making predictions on new episodes, the model is essentially trained on episodes that were earlier in date/time than the new episodes. Consequently, by randomly sampling a test set, this consideration is neglected because there is no date/time distinction made between the episodes in the train and test sets. 5-fold cross-validation was performed on the train set for model and hyperparameter selection. The train set was shuffled during cross-validation and a random state was set. To prevent unfair shrinking of coefficients for the regularized linear models (Models 2-4) and prevent the domination of large values in the Support Vector Regression model, the data was scaled and centered before training these models.

Model tuning was performed differently depending on the models, for the unregularized and regularized linear models (Models 1-4), a grid of values was searched over multiple times, manually setting the parameter values each time, until the best parameters were found. This method was chosen because the number of parameters to tune is quite small for these algorithms, making it computationally feasible to search over a whole grid. For models 5-7, the parameters were selected based on randomly picking a set of parameters based on a parameter space, and after a set number of iterations, picking the best parameters. This method was chosen because it is not computationally feasible to exhaustively search over a grid because the number of parameters for the algorithms is higher. Finally, because the tree-based algorithms (Models 6 and 7) showed significant overfitting, after the parameter search, key complexity parameters such as the maximum tree depth were reduced up to a point where significant decreases in cross-validation R^2 started to occur ($\Delta R^2 > 0.05$). For the regularized linear regression models (Models 2-4), alpha was tuned (which represents either the L1 or L2 regularization strength) and for the Elastic Net Regression model, the l1_ratio was also tuned. For the Support Vector Regression model (5), C, gamma and kernel were tuned. For the Random Forest Regression model (6) the hyperparameters n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf and bootstrap were tuned. For the Extreme Gradient Boosting Regression model (7) the hyperparameters learning_rate, n_estimators, max_depth, subsample, colsample_bytree, gamma and n_estimators were tuned.

Furthermore, to answer the second question, the permutation feature importance method was used. The permutation feature importance method calculates the increase in the prediction error of the model after the values of a feature are permuted, which removes the

relationship between the feature and the target (Molnar, 2020). Hence, the value of feature importance implies the increase in prediction error, and the larger the increase in prediction error, the more important the feature. The permutation feature importance method is a model-agnostic method, meaning it can be used for any model. Hence, the flexibility to choose any model is retained, no matter its complexity. Furthermore, the feature importance method considers all interactions, not only with the target variable but also with other predictor variables, because their interaction is also removed due to the permutation (Molnar, 2020). This method was chosen because of its clear interpretation, flexibility, and ability to account for all interactions.

Topic Modeling and Topic Feature Analysis

The second aim of this thesis is to investigate whether useful features can be extracted using topic modeling on the subtitles of episodes in the context of public broadcasting and whether topics be identified that are likely to generate high online participation. Whether a feature or set of features is useful in a prediction depends on its impact on the performance of a model. Furthermore, as discussed earlier, the distinction in the importance of features depends on how much they contribute to the prediction of the target. Thus, translating the research questions to data science questions, this thesis aims to answer (3) '*Is there a significant improvement in R^2 and RMSE when including the features extracted from topic modeling?*' and (4) *'Which topics contribute most to the prediction of online participation?'*.

To preprocess the subtitles, they were lemmatized using spaCy (version 2.3.5). The subtitles were lemmatized rather than stemmed to obtain the meaningful base forms of words instead of the often much less meaningful stems. As the meaning of the words must be interpreted to determine what topics the subtitles are about, lemmatization is the preferred method of preprocessing for answering the research questions. Punctuations, white spaces, numbers have no relevant meaning in this context and therefore only create noise, hence, they were excluded in the lemmatization process. Also, capitalized words do not provide relevant additional meaning as opposed to lower case words in this context, hence, all capitalized words were transformed to lower case words. Furthermore, Dutch stop words and a couple of meaningless words that showed up in the first run of the topic model were excluded, words such as 'you', 'no' and 'hi'. These were excluded from the topic model because they are not informative of a certain topic, and therefore only contribute noise to the output of the topic model.

Subsequently, an LDA topic model was trained using the Gensim library (version 4.0.1) and MALLET (version 2.0.8). A Latent Dirichlet Allocation (LDA) topic model is a statistical

model that aims to find latent topics of large amounts of text efficiently while preserving the essential statistical information (Blei et al., 2003). In an LDA topic model, each document (i.e. episode subtitles in this case) can be described by a distribution of topics and each topic can be described by a distribution of words, both modelled as Dirichlet distributions. LDA tries to find the two Dirichlet distributions that best represent the original documents by starting with random assignments of words to topics and iteratively updating these. LDA iterates over all words in all documents and re-assigns the words to topics according to two considerations. Firstly, LDA considers how many times the respective word occurs in the respective document; if a lot of the words in the document belong to a topic it becomes more likely that the respective word will be assigned to that topic. Secondly, it considers how many times the respective word is assigned to a certain topic over all documents; if the word is often assigned to a certain topic, the respective word will likely also be assigned to it. At the end of each iteration, the respective word is re-assigned to the topic that scores best based on these two considerations. When creating the dictionary, the minimum document frequency was set to 10 and the maximum document frequency to 0.70. The numbers of topics 30, 50, 100 and 200 were tried and the number of topics that qualitatively created the most meaningful topics was chosen, which was 50 topics. The number of iterations for the topic model was set to 1000.

As discussed earlier, the topic model was trained on a larger dataset than that is used for answering the research questions. The topic distributions, containing one feature per topic, are used as features in the prediction of online participation and are between 0 and 1. The topic features were checked for multicollinearity, and no Pearson correlation coefficients exceeded 0.80. The topic features were included in all models previously discussed (Models 1-7) and were tuned again following the same process described earlier in this section to make sure the approximate full potential of these models is reached. To answer whether there is a significant increase in performance when including these topic distributions as features, the R^2 and *RMSE* on the test set of the best performing model without topic features were compared to the R^2 and *RMSE* on the test set of the best performing model with topics. To better interpret the *RMSE*, the predictions and actual values on the test set were back-transformed by taking $y' = e^y + 1$, where y represent the log-transformed predictions or actual values of online participation on the test set.

To answer which topics contribute most to the prediction of online participation, the permutation feature importance method, as discussed earlier, is used to determine whether there are topics that contribute highly to the prediction of online participation and which these are.

When considering the earlier introduced features that imply the categories the episode belongs to (*SKO Class L1, SKO Class L2, CCC Domain Definition, CCC Definition)*, it should be noted that these features are in a way similar to the topics that are generated from topic modeling, because they imply the category of content of the episode. However, we can expect that topic modeling does contribute to the prediction in additional ways. Firstly, the topics can be more specific than the category, which is designed to be relatively broad. For example, a topic can be ‘soccer’ while ‘soccer’ is likely too specific to be a category. Secondly, as topic distributions are included as features, this allows episodes to belong to multiple topics, while this is not allowed for the categories, as an episode only belongs to one category. As episodes can be about different topics, the topic features likely bring additional information to the model that may be useful. These differences between categories and topics are driven by the way topic modeling operates, which has been discussed earlier in this section.

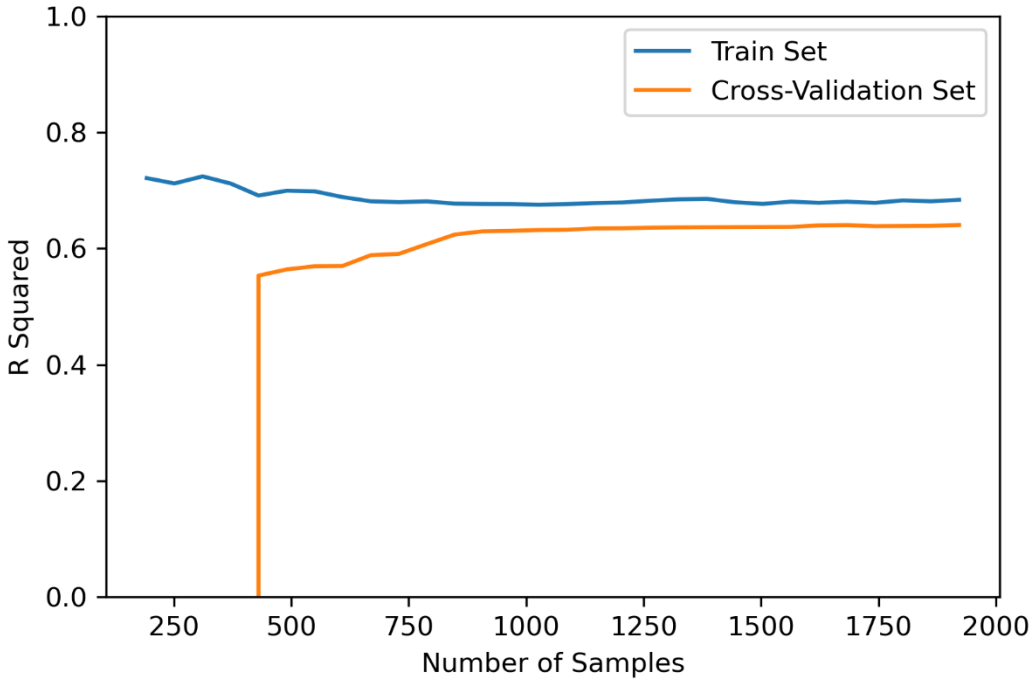
RESULTS

Table 1. Prediction Results.

	1. Linear	2. Ridge	3. LASSO	4. Elastic Net	5. Support Vector	6. Random Forest	7. Extreme Gradient Boosting
Excl. Topics							
Train R^2	0.681	0.678	0.676	0.676	0.670	0.793	0.79
CV R^2	0.641	0.647	0.645	0.647	0.648	0.675	0.667
Train *RMSE*	0.993	0.998	1.000	1.000	1.010	0.799	0.806
CV *RMSE*	1.052	1.043	1.046	1.044	1.041	1.000	1.012
Test R^2						0.654	
Test *RMSE*						0.940	
Test *RMSE* (Log-reversed)						453	
Incl. Topics							
Train R^2	0.698	0.691	0.687	0.684	0.686	0.974	0.905
CV R^2	0.641	0.651	0.65	0.65	0.647	0.672	0.661
Train *RMSE*	0.966	0.978	0.983	0.988	0.985	0.281	0.543
CV *RMSE*	1.052	1.038	1.039	1.038	1.043	1.005	1.022
Test R^2						0.663	
Test *RMSE*						0.926	
Test *RMSE* (Log-reversed)						430	

The prediction results are presented in Table 1. Table 1 presents two scores for model evaluation and selection: R^2 and *RMSE*, and the results are split between models with the topic features and without. It has to be noted that the *RMSE* is equal to the *RMSLE* because the target has been transformed by taking $y' = \ln(y) + 1$, where y represents online participation. For the purpose of model selection, in the Table, the presented scores for all models are the train set scores and the mean cross-validation scores. As it is not desirable to choose a model based on the test set, because information about the test set may leak into training, the test results are only presented for the highest performing models, which are the models that are chosen based on cross-validation scores. This means the test scores can still be compared between the best model with topic features and the best model without topic features. For the test scores, the *RMSE* is also presented with the reversion of the log transformation of online participation on the actual test set and the prediction on the test set.

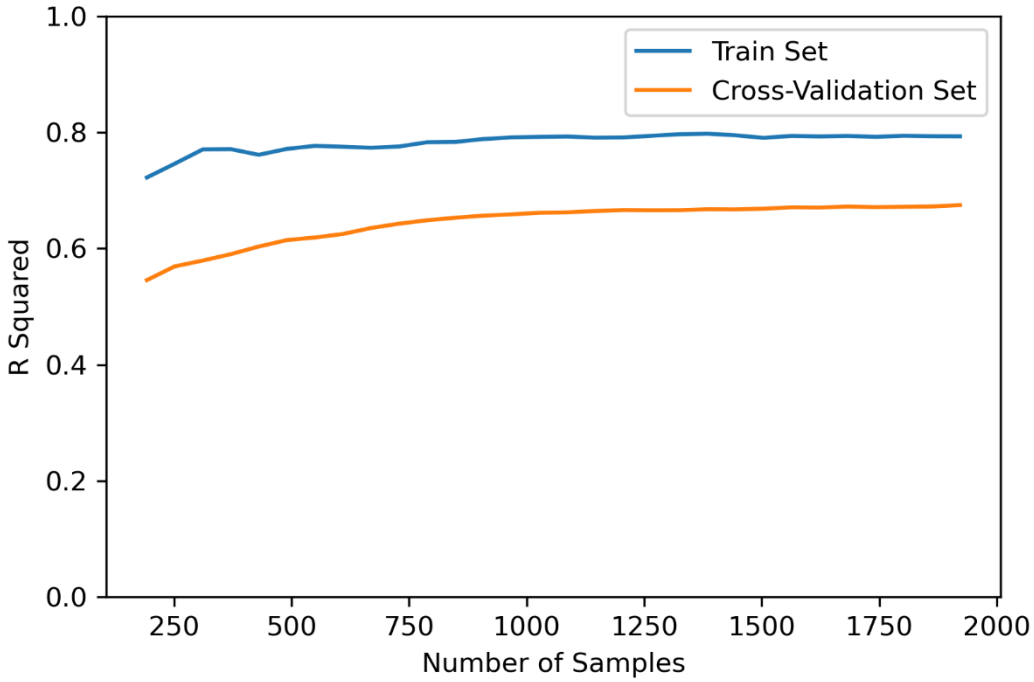
Figure 3. Learning Curve of Linear Regression (1) Excl. Topics



The performance of models is compared in terms of cross-validation scores. All main model hyperparameters were tuned before evaluation. Shown in Table 1, the cross-validation results for the linear regression model without regularization and excluding topics (1) are decent for the simplest model in the set of models, where 64.1% of the variance in the of (log-

transformed) online participation is explained by the model (R^2 = 0.641). As shown in Figure 3, the learning curve of this model illustrates that the model is slightly overfitting. The regularized linear models (Models 2-4) do not introduce much difference (0.645 < R^2 < 0.647) compared to the non-regularized linear model. This is expected when the non-regularized linear model is not overfitting much, as the goal of regularization is to reduce overfitting. However, the regularized linear models do not manage to significantly reduce the slight gap between the train and cv scores, as the increase in R^2 is minimal compared to the unregularized linear model. The support vector regression model does show similar performance to the other linear models (R^2 = 0.648).

Figure 4. Learning Curve of Random Forest Regression (6) Excl. Topics



As shown in Table 1, the results of the tree-based models (Models 6 and 7) indicate that these models overfit the data significantly more than the previous models. This is also illustrated by the learning curve of the random forest model, shown in Figure 4. Nevertheless, cross-validation scores do improve slightly, with the random forest model achieving the best performance (R^2 = 0.675) of all models. It must be noted that the problem of overfitting is controlled by the hyperparameters set in the model. During model tuning, the best hyperparameters were chosen based on cross-validation performance, and subsequently, to

reduce overfitting, key complexity hyperparameters such as maximum tree depth was decreased until significant decreases in cross-validation scores occurred ($\Delta R^2 > 0.05$). Hence, increased overfitting was accepted only when significant increases in cross-validation performance could be achieved, which is something to consider when interpreting the results. All in all, increasing model complexity and adding regularization did only slightly increase model performance over the simple, unregularized linear regression model while the more complex, tree-based models introduced significantly more overfitting. Furthermore, as indicated by both Figure 3 and 4, increasing the number of samples (x-axis) beyond the current number of samples does not seem to be a fruitful way to increase performance, if we were to extrapolate the line further beyond the right side of the plot.

Table 2. Topic Labels and Word Topic Distributions

Topic	**Label**	**Word Topic Distributions**
1	Corona Measures	corona 1.51%, maatregelen 1.05%, open 0.93%, coronacrisis 0.9%, kabinet 0.79%, houden 0.77%, horeca 0.66%, mogen 0.63%, afstand 0.63%, ondernemer 0.6%
2	Corona Vaccine	corona 1.43%, vaccin 1.15%, virus 1.12%, aantal 0.94%, maatregelen 0.87%, testen 0.83%, week 0.73%, ziekenhuizen 0.67%, besmettingen 0.58%, ic 0.58%
3	Locations	water 2.81%, zee 1.14%, land 0.81%, schip 0.78%, ligt 0.73%, lang 0.71%, weg 0.59%, eiland 0.58%, meter 0.57%, stad 0.57%
4	Christianity	god 3.28%, kerk 2.47%, onze 1.88%, jezus 1.47%, leven 1.36%, heer 1.22%, geloof 0.96%, bidden 0.72%, christus 0.69%, heilige 0.68%
5	Dutch Soccer International	wedstrijd 1.32%, ajax 1.14%, oranje 1.05%, club 1.01%, speler 0.92%, league 0.9%, spelen 0.87%, voetbal 0.76%, finale 0.73%, bal 0.64%
6	Family	moeder 3.78%, vader 3.41%, kinderen 1.87%, kind 1.35%, familie 1.12%, leven 1.11%, ouder 0.89%, dochter 0.78%, elkaar 0.78%, ouders 0.74%
7	Government	minister 1.86%, kamer 1.14%, kabinet 1.02%, vraag 0.92%, vragen 0.67%, tweede 0.58%, heer 0.52%, voorzitter 0.52%, debat 0.46%, dank 0.46%
8	Justice	rechter 1.39%, zaak 1.24%, politie 0.96%, advocaat 0.84%, rechtbank 0.78%, justitie 0.71%, onderzoek 0.71%, moord 0.67%, zegt 0.58%, openbaar 0.56%
9	Sustainability	ziet 1.02%, zit 0.88%, onze 0.82%, soort 0.65%, plastic 0.64%, kijken 0.59%, wereld 0.58%, bijvoorbeeld 0.58%, kun 0.57%, aarde 0.55%
10	Living	huis 2.24%, mooi 1.39%, wonen 0.95%, helemaal 0.69%, kijk 0.67%, kamer 0.65%, tuin 0.62%, oude 0.61%, zie 0.54%, allemaal 0.54%
11	Tour De France	tour 0.98%, rijden 0.79%, zit 0.78%, ploeg 0.64%, etappe 0.62%, kilometer 0.59%, man 0.56%, laatste 0.51%, rijdt 0.51%, peloton 0.48%
12	Royal Family	koningin 1.58%, willem 1.44%, koning 1.41%, kone 1.17%, alexander 0.84%, maxima 0.75%, koninklijke 0.7%, prinse 0.63%, prin 0.63%, familie 0.6%
13	Meaningless 1	keer 4.03%, terug 2.58%, vier 2.53%, laatste 2.39%, drie 2.06%, stappen 2.01%, voren 1.97%, omhoog 1.57%, plaat 1.33%, kom 1.3%
14	Food	mooi 1.35%, eten 1.07%, gemaakt 0.69%, doe 0.68%, zit 0.58%, ziet 0.58%, kijk 0.57%, mooie 0.56%, smaak 0.53%, zie 0.51%
15	Meaningless 2	dank 1.37%, denk 1.35%, doe 0.85%, weet 0.84%, hallo 0.84%, haha 0.83%, zeker 0.77%, keer 0.74%, leuke 0.68%, erg 0.68%

16	Education	kinderen 4.82%, school 3.54%, scholen 2.0%, onderwijs 1.57%, leerlingen 1.49%, jongeren 1.13%, studenten 1.03%, krijgen 0.98%, ouders 0.96%, ouder 0.82%
17	Illness	ziekte 0.83%, ziekenhuis 0.78%, leven 0.65%, onderzoek 0.65%, ziek 0.6%, krijgen 0.6%, lichaam 0.53%, medicijnen 0.53%, krijgt 0.51%, vaak 0.51%
18	Meaningless 3	jou 2.84%, m'n 1.39%, weet 1.27%, laat 1.2%, leven 1.17%, nooit 1.0%, bent 0.93%, liefde 0.81%, la 0.81%, week 0.66%
19	Gender	vrouwen 1.52%, zeggen 0.95%, mannen 0.74%, man 0.72%, vrouw 0.64%, soort 0.58%, elkaar 0.57%, denk 0.56%, zegt 0.55%, waarom 0.52%
20	Media Consumption	boek 1.09%, hebt 0.94%, programma 0.9%, kijken 0.86%, film 0.63%, zit 0.61%, iet 0.6%, jou 0.48%, tv 0.47%, vond 0.46%
21	Ice Skating	meter 1.38%, laatste 0.75%, rijden 0.73%, seizoen 0.57%, keer 0.54%, goede 0.52%, schaatsen 0.51%, tweede 0.5%, wk 0.5%, record 0.5%
22	Weather	graden 2.38%, zon 1.26%, morgen 1.25%, uur 1.07%, wind 0.97%, regen 0.92%, droog 0.89%, volgens 0.85%, buien 0.77%, noorden 0.65%
23	Sports	finale 0.83%, tweede 0.79%, spelen 0.77%, wedstrijd 0.75%, set 0.61%, sport 0.54%, bal 0.52%, olympische 0.52%, laatste 0.46%, keer 0.46%
24	Meaningless 4	kom 1.63%, weet 1.56%, gelach 0.99%, zeg 0.9%, hoor 0.87%, kijk 0.82%, niks 0.79%, doe 0.79%, muziek 0.78%, hallo 0.7%
25	Meaningless 5	stap 1.67%, antwoord 1.54%, laatste 1.22%, max 1.13%, volgende 1.07%, dank 0.94%, inderdaad 0.91%, hoor 0.82%, vraag 0.8%, bent 0.77%
26	International Relations	rusland 0.99%, turkije 0.88%, nederlandse 0.85%, land 0.83%, landen 0.82%, iran 0.8%, president 0.72%, amerikaanse 0.65%, noord 0.61%, europa 0.56%
27	Meaningless 6	denk 1.95%, zeggen 1.42%, misschien 1.27%, kijken 1.26%, natuurlijk 1.16%, zit 1.15%, allemaal 1.13%, zitten 0.96%, hebt 0.93%, ander 0.88%
28	Meaningless 7	denk 1.8%, leven 1.66%, dingen 1.13%, iet 1.06%, hebt 1.02%, jou 0.94%, weet 0.93%, bent 0.93%, moment 0.76%, erg 0.72%
29	Finance	euro 2.16%, geld 1.97%, betalen 0.76%, krijgen 0.57%, bank 0.53%, bedrijf 0.52%, onze 0.51%, krijgt 0.44%, zegt 0.42%, terug 0.41%
30	Meaningless 8	kwam 1.84%, hadden 1.79%, moment 1.35%, dacht 1.29%, moest 1.28%, zat 1.15%, zag 1.05%, nooit 1.0%, vond 0.98%, wilde 0.97%
31	Meaningless 9	weet 1.87%, niks 0.95%, waarom 0.92%, jou 0.89%, school 0.86%, misschien 0.83%, iet 0.81%, doe 0.74%, hebt 0.69%, ander 0.67%
32	Business	geld 1.7%, bedrijven 1.1%, willen 0.82%, procent 0.7%, minder 0.67%, betalen 0.66%, economie 0.61%, euro 0.58%, miljard 0.54%, miljoen 0.54%
33	Dutch Politics	partij 1.66%, rutte 1.23%, partijen 1.11%, vvd 1.09%, cda 1.0%, kamer 0.99%, politiek 0.9%, goedemorgen 0.77%, tweede 0.75%, verkiezingen 0.72%
34	War	oorlog 1.42%, land 0.63%, werden 0.63%, vrijheid 0.59%, duitse 0.58%, wereldoorlog 0.53%, leven 0.52%, nederlandse 0.52%, tweede 0.48%, duitsers 0.48%
35	American Politics	trump 3.06%, president 1.99%, biden 1.15%, brexit 1.02%, europese 0.98%, verkiezingen 0.96%, amerikaanse 0.8%, eu 0.76%, amerika 0.74%, premier 0.73%
36	China	china 1.99%, chinese 0.98%, schiphol 0.77%, volgens 0.64%, israël 0.5%, vliegtuig 0.47%, aantal 0.47%, uur 0.45%, graden 0.41%, coronavirus 0.4%
37	Protest	politie 3.11%, burgemeester 1.23%, demonstranten 0.81%, demonstratie 0.79%, geweld 0.72%, stad 0.71%, agenten 0.67%, straat 0.65%, amsterdam 0.59%, rotterdam 0.54%
38	Team Game Show	vraag 2.0%, team 1.47%, applaus 1.03%, spelen 1.02%, antwoord 1.01%, volgende 0.92%, drie 0.91%, punten 0.9%, punt 0.85%, vragen 0.77%
39	Quiz Show	pas 1.89%, vraag 1.87%, heet 1.08%, weet 1.02%, mag 1.0%, seconden 0.94%, ronde 0.93%, meneer 0.89%, welke 0.87%, mevrouw 0.74%
40	Music	muziek 3.34%, nummer 1.5%, zingen 0.85%, love 0.77%, liedje 0.76%, applaus 0.72%, spelen 0.59%, songfestival 0.43%, horen 0.43%, zingt 0.42%

41	Art	mooi 1.15%, gemaakt 0.89%, museum 0.84%, eeuw 0.72%, werk 0.7%, schilderij 0.7%, zie 0.68%, kunst 0.61%, natuurlijk 0.58%, ziet 0.56%
42	Meaningless 10	kom 1.02%, kijk 0.88%, denk 0.87%, hebt 0.79%, jongens 0.77%, zit 0.72%, klaar 0.72%, muziek 0.68%, kijken 0.65%, doe 0.65%
43	News	graden 0.93%, volgens 0.91%, goedemorgen 0.84%, onderzoek 0.7%, euro 0.68%, procent 0.65%, miljoen 0.65%, vorig 0.64%, blijkt 0.6%, ruim 0.6%
44	Dutch Soccer National	bal 1.6%, ajax 1.4%, psv 1.26%, feyenoord 1.06%, wedstrijd 1.05%, az 0.79%, seizoen 0.58%, goal 0.56%, fc 0.54%, helft 0.53%
45	Car Accident	auto 1.99%, man 1.68%, politie 0.95%, dader 0.66%, weten 0.63%, mannen 0.6%, slachtoffer 0.59%, mogelijk 0.58%, vrouw 0.51%, iet 0.51%
46	Healthcare	zorg 1.58%, nodig 1.12%, werk 1.01%, elkaar 0.99%, helpen 0.95%, thuis 0.86%, zorgen 0.83%, hulp 0.82%, ouderen 0.74%, werken 0.69%
47	Children	kinderen 2.59%, vinden 0.64%, bijvoorbeeld 0.6%, daarom 0.51%, erg 0.5%, nieuws 0.49%, land 0.48%, iedereen 0.46%, jeugdjournaal 0.45%, krijgen 0.43%
48	Fires	brand 1.95%, politie 1.07%, uur 1.0%, vuurwerk 0.76%, volgens 0.75%, brandweer 0.65%, vuur 0.6%, afgelopen 0.59%, regionale 0.51%, omroepen 0.5%
49	Animals	dieren 1.79%, hond 0.71%, natuur 0.69%, zie 0.66%, vogel 0.61%, dier 0.61%, zit 0.6%, eten 0.56%, zitten 0.55%, bos 0.55%
50	Farmers Protest	boeren 2.0%, den 0.91%, haag 0.83%, willen 0.75%, provincie 0.6%, groningen 0.51%, weg 0.5%, stikstof 0.49%, bedrijf 0.48%, nieuwe 0.48%

The topic labels and respective word topic distributions generated by the topic model are presented in Table 2. The topic features included in the model are the document topic distributions generated by the topic model, with one feature per topic with values ranging between 0 and 1. The model hyperparameters differ between the models with topic features and without because they were tuned separately. As shown in Table 1, the cross-validation results of the models with topic features are very similar to the results of models with topic features. Again, the random forest model performs best out of all models (R^2 = 0.672). However, the scores on the train set are noticeably higher for the tree-based models, which indicates the models are overfitting the data more than the models without topic features. However, as discussed earlier, the hyperparameters control the extent of overfitting of the model and the parameters were tuned separately for the models with topic features. For the models with topic features, higher cross-validation could be attained by increasing model complexity hyperparameters such as maximum tree depth. Thus, this likely explains the difference in overfitting and the extra overfitting should probably not be considered a direct result of the characteristics of the topic features, but rather a result of increasing the number of features and how that affected the extent of achievable model performance through tuning of complexity parameters.

For both the models with topic features and without, the random forest model performed best in terms of cross-validation R^2 and *RMSE* scores. Hence, now that the best

models have been identified based on cross-validation results, their test scores can be compared to see what impact the inclusion of topic features has on the prediction of online participation. The random forest model without topic features explains 65.4% of the variance in (log-transformed) online participation (R^2 = 0.654) and its *RMSE* based on the back-transformed prediction is approximately 453. This means that the prediction of online participation is, on average, different from the true value by 453 comments. The random forest model with topic features explains 66.3% of the variance in (log-transformed) online participation (R^2 = 0.663), and its *RMSE* based on the back-transformed prediction is approximately 430. This means that the prediction of online participation is, on average, different from the true value by 430 comments. This is an increase of 1.4% in terms of R^2 and a decrease of 5.1% in terms of *RMSE* based on back-transformed predictions. Although these slight performance improvements were found on the test set, when comparing the cross-validation scores the performance is slightly worse when the topic features are included, therefore questioning the validity of these results. Considering this, it is likely that including the topics has very little effect on the prediction performance of the model.

Figure 5. True versus Prediction for Random Forest Excl. Topic Features

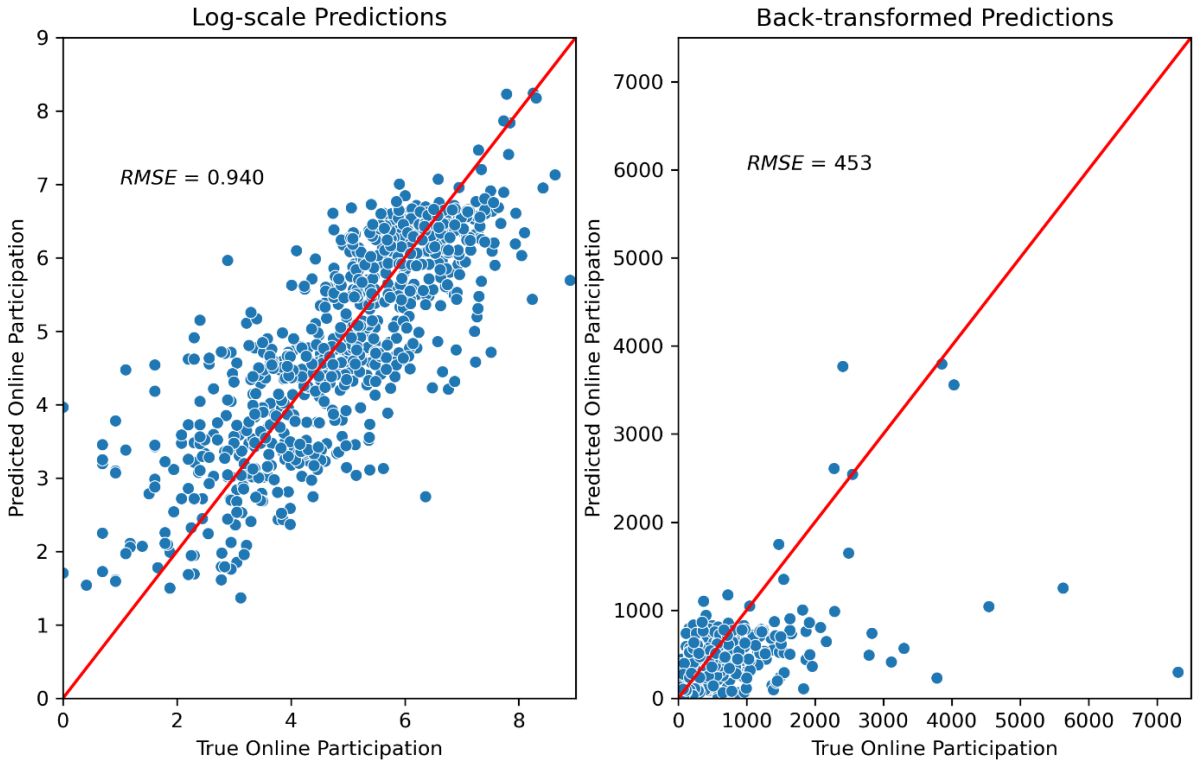
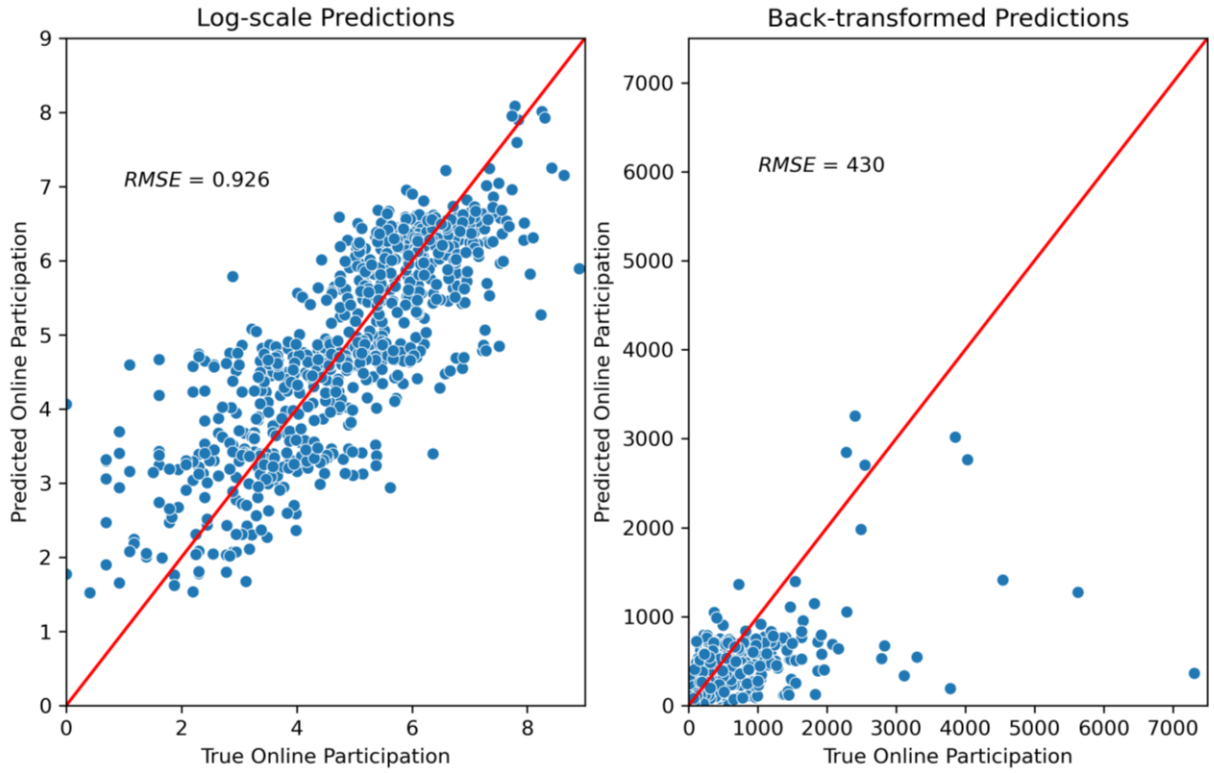


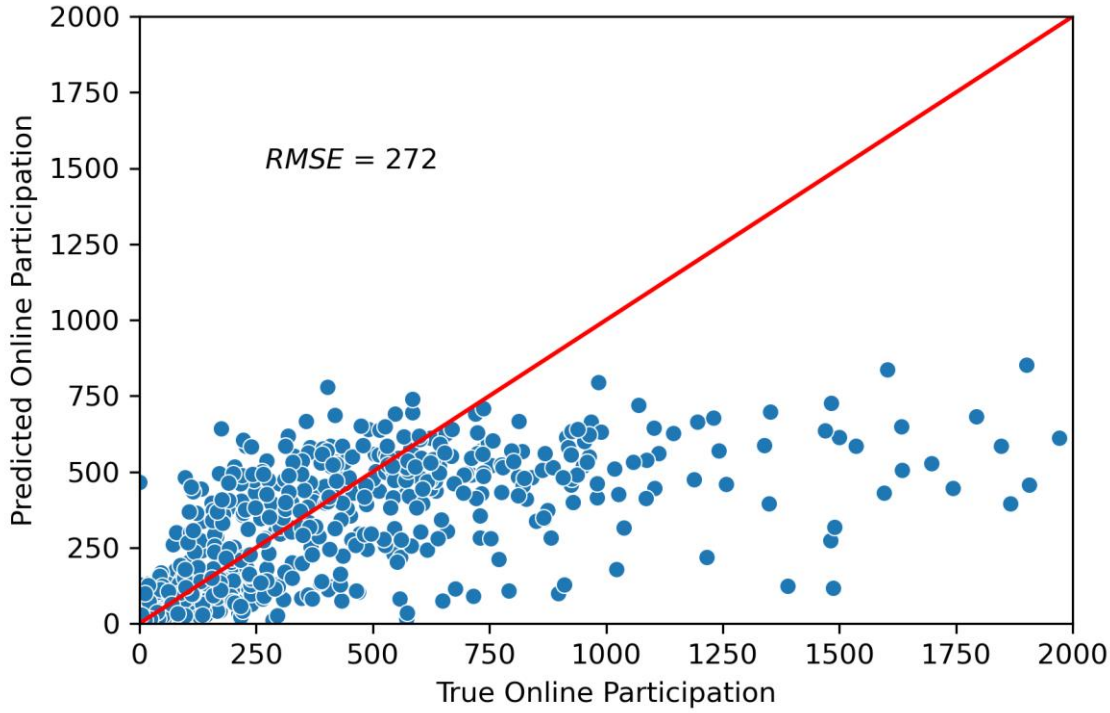
Figure 6. True versus Prediction for Random Forest Incl. Topic Features



Figures 5 and 6 illustrate the predictions on the test set (y-axis) versus the true values on the test set (x-axis), for the random forest model without topic features (Figure 5) and with topic features (Figure 6). In these Figures, the values are presented on the log scale on which the model was trained (left) and the back-transformed scale (right). The line drawn through the middle represents the points at which the prediction is equal to the true value. Both Figures show similar results, further indicating that the models perform very similar. On the log-transformed scale, the predictions are decent, which is a result of the bringing the very large outliers closer to the bulk of the distribution by performing the log-transformation. On the back-transformed scale, however, the error increases as online participation increases. Even though online participation was log-transformed (and back-transformed to interpret the results), on the back-transformed scale, the model still seems to have trouble predicting the higher values of online participation, where the values on the right are very high but are predicted to be much lower. For example, the value on the far right represents a true online participation of around 7000, while the model predicted it to be around 500. These false predictions likely have a large impact on the average error and indicate a problem of heteroscedasticity, which raises the

question of how this model would score on average without these outliers, or rather with a limited set of episodes that are limited in online participation.

Figure 7. True versus Prediction for Random Forest Incl. Topic Features for Values < 2000



Removing all values above 2000 before training the random forest model with topic features resulted in a decrease of 89 episodes, or 2.8% of the total episodes. With an R^2 of 0.662, the explained variance in (log-transformed) online participation remained approximately equal, which makes sense because the changes are not expected to make significant changes on the log scale because the values are much closer together on that scale. However, as expected, the *RMSE* based on back-transformed predictions, dropped significantly to a value of approximately 272. This means that the model's prediction of online participation is, on average, different from the true value by 272 comments. This is a decrease of 36,7% compared to the *RMSE* of the model with all values of online participation, meaning the model is 36,7% more accurate while retaining 97.2% of the data. Figure 7 represents the new true versus prediction Figure for this model on the back-transformed scale. As shown in the Figure, by excluding the few very high values, the significant gaps between true and predicted online participation are reduced, resulting in the lower *RMSE*. Hence, the predictive capability of the random forest models discussed so far is significantly higher for lower values of online

participation. This is likely because as values of online participation increase, the data becomes fewer, making it harder for the algorithm to learn what these values should be, even through the use of a log transformation.

As shown in Figures 5 and 6, although the errors on the back-transformed scale indicate high homoscedasticity, for the predictions on the log-scale, the errors seem to be approximately heteroscedastic. Therefore, the *RSME* on the back-transformed scale is meaningless for many values of online participation. For example, for lower values of online participation, the average error is much lower than the *RMSE* of all values, which is 430 comments. Hence, this *RMSE* is uninformative of the error we can expect for those values, and this is a result of the extremely high values of online participation that skew the *RMSE*. However, the *RMSE* on the log-scale is informative, as the line in Figures 5 and 6 separates the errors nicely, meaning an *RMSE* on the log-scale of 0.926 is a good approximation of the error at most values of online participation. The *RMSE* on the log-scale of 0.926 means an average difference by a factor of $e^{0.926} \approx 2.52$ from the actual value of online participation. If the model would predict a value on the log-scale of 2, this would mean a real-value prediction of $e^2 - 1 \approx 6.39$, and on average, we expect that this is accurate within a factor of 2.52, which presents a relatively low real-value error. For higher values of online participation, the multiplicative interpretation holds but the real-value interpretation is different. For example, a prediction on the log-scale of 8, would result in a real-value prediction of $e^8 - 1 \approx 2979.95$, and now expecting that this is accurate on average within a factor of 2.52, the real-value error becomes much higher when compared to lower values of online participation. However, in multiplicate terms, the interpretation holds. Thus, in multiplicative terms, the average error is informative as the average distance by a factor of 2.52 can be interpreted for most values as opposed to the *RMSE* of 430. This also indicates that the model performs quite well on lower values of online participation because an average error by a factor of 2.52 is reasonably small for lower values.

Table 3. Top-20 Important Features of Random Forest Model Excl. Topics

Feature	Imp.	Corr.
Own Posts Comments	0.080	0.503
Page Views Total	0.072	0.534
Historical Average Views	0.058	0.504
Fans	0.041	0.250
Duration	0.038	0.213
Page Engagement Rate Unique	0.035	0.348
Fans Change	0.035	0.296
Own Posts	0.034	0.391
Page Impressions Paid Unique	0.032	0.375
Channel: NPO 2	0.025	0.546
Own Posts Engagement Rate	0.021	0.009
First Broadcaster: AVTR	0.013	0.483
Channel: NPO 1	0.012	0.546
SKO Class L2: Other Non-Fiction	0.011	0.399
Title: Binnenstebuiten	0.009	0.343
NPO CCC Domain Definition: Opinion	0.009	0.360
SKO Class L2: News & Actualities	0.008	0.401
First Broadcaster: KRNC	0.008	0.288
NPO CCC Domain Definition: Knowledge	0.007	0.286
Title: Eenvandaag	0.007	0.441

Table 4. Top-20 Important Features of Random Forest Model Incl. Topics

Feature	Imp.	Corr.
Own Posts Comments	0.083	0.503
Page Views Total	0.060	0.534
Historical Average Views	0.048	0.504
Fans	0.030	0.250
Page Engagement Rate Unique	0.030	0.348
Duration	0.025	0.213
Channel: NPO 2	0.025	0.546
Fans Change	0.024	0.296
Page Impressions Paid Unique	0.021	0.375
Own Posts	0.020	0.391
First Broadcaster: AVTR	0.014	0.483
Channel: NPO 1	0.014	0.546
Own Posts Engagement Rate	0.011	0.009
SKO Class L2: Other Non-Fiction	0.011	0.399
Topic 10: Tour de France	0.010	0.339
Topic 29: Meaningless 8	0.009	0.253
Topic 21: Weather	0.008	0.204
Title: Binnenstebuiten	0.008	0.343
NPO CCC Domain Definition: Opinion	0.008	0.360
Title: Eenvandaag	0.008	0.441

In Tables 3 and 4, the top-20 important features of the random forest models with topic features (left) and without topic features (right) are presented. It has to be noted that the feature importances only indicate the increase in model error when the feature is permuted, and neither can distinctions be made in terms of whether it positively or negatively affects online participation itself nor can exact effect sizes be interpreted. To give an indication of the linear bivariate relationship between the feature and the target, the Pearson correlation coefficients have also been included in the Tables. This is by no means a rigorous metric to identify the effects between each of the features on online participation in this multivariate context, however, these correlations do give an indication of the bivariate relationship, which assists in giving an indication of the direction of the effect. All in all, the two Tables are very similar. For both models, the average historical social media engagement features *Own Posts Comments*, *Page Views Total*, *Fans*, *Fans Change*, *Page Impressions Paid Unique* and *Own Posts Engagement Rate* are important, with *Own Posts Comments* being the most important one in both models, which is not surprising as online participation is measured through comments as well. *Historical Average Views* is also important in both models, as was anticipated earlier. Because correlations for both the historic average social media engagement features and the *Historical Average Views* are positive, the results suggest that, on average, higher social media engagement or *Historical Average Views*, is associated with higher online participation.

After the top-10 important features, episode characteristics start to show importance such as specific categories for *First Broadcaster, Channel, Duration* and *Title,* and specific categories for the features that categorize an episode such as *NPO CCC Domain Definition* and *SKO Class L2*. Specifically, whether an episode is broadcasted on the channel NPO 2 is important in predicting online participation, indicating that episodes on that channel are on average, associated with higher online participation than other channels such as NPO 1, which shows up later in the list. Also, two titles show up in both top-20s: Binnenstebuiten and Eenvandaag, and several categories: Other Non-Fiction, Opinion, News & Actualities and Knowledge, indicating that, due to the positive correlations, episodes that belong to these shows and categories are on average, associated with higher online participation. After the top-15, some more interesting changes start to occur, where three topic features take the place of episode characteristic features, one of which is meaningless. Nonetheless, the topics Weather and Tour de France show to be relatively important in the model considering there are 168 features in total.

Table 5. Top-10 Most Important Topics

Feature	Imp.	Corr.
Topic 10: Tour de France	0.010	0.339
Topic 29: Meaningless 8	0.009	0.253
Topic 21: Weather	0.008	0.204
Topic 31: Business	0.007	0.149
Topic 43: Dutch Soccer National	0.007	0.163
Topic 24 Meaningless 5	0.007	0.069
Topic 19: Media Consumption	0.007	0.217
Topic 4: Dutch Soccer International	0.006	0.188
Topic 46: Children	0.006	0.027
Topic34: American Politics	0.006	0.156

Table 6. Top-10 Least Important Topics

Feature	Imp.	Corr.
Topic 8: Sustainability	0.003	0.040
Topic 32: Politics	0.003	0.081
Topic 14: Meaningless 2	0.003	0.068
Topic 3: Christianity	0.003	0.044
Topic 26: Meaningless 6	0.003	0.124
Topic 41: Meaningless 10	0.003	0.000
Topic 7: Justice	0.003	0.096
Topic 6: Government	0.003	0.028
Topic 13: Food	0.003	0.080
Topic 36: protest	0.003	0.085

Tables 5 and 6 present the top-10 most important topics (left) and the top-10 least important topics (right). The overall increase in model performance due to the additional topic features is small. Hence, although distinctions between topics can be made, as expected, no individual topic has a large effect on the prediction error of online participation and the differences in the impact on the prediction are small between topics. The topics Tour de France, Weather, Business, Dutch Soccer National and International, Media Consumption, Children and American Politics are the most important topics when predicting online participation. Furthermore, the least important topics are Sustainability, Politics, Christianity, Justice, Government, Food and Protest. Interestingly, American Politics shows up in the most important topics while Politics, which is about politics in the Netherlands, shows up in the least important topics. Due to the positive correlations, this indicates that episodes about American Politics generate more online participation than episodes on Dutch Politics.

DISCUSSION

Answers to the Research Questions

Returning to the initial research questions, firstly, this thesis aimed to answer to what extent online participation could be predicted in the context of public broadcasting, through investigating what performance could be achieved on a predictive regression model with online participation as a target. To investigate this, 22 potentially predictive features of online participation were collected and created, missing values and outliers were dealt with, and 7 models were individually tuned and compared with the aim of achieving the best R^2 and *RMSE*. Of these models, the random forest model performed best with an R^2 of 0.675 and an *RMSE* based on back-transformed predictions of 453. Hence, online participation in the context of public broadcasting online participation can be predicted with an average error of 453 comments. Visualizations of the results indicated that there were large errors for high values of online participation, even though a log-transformation was applied. This was confirmed by running the model only with lower values of online participation, which resulted in a significantly lower *RMSE* based on back-transformed predictions of 272. Hence, the model struggles with higher values of online participation, likely because the data becomes fewer as online participation increases. Furthermore, learning curves of the models where the performance of the models is plotted against the number of samples indicated that increasing the number of samples would likely not be a fruitful way to increase performance. Although, this might be the case if the samples were to reduce the long tail of online participation, which would likely reduce the high errors that are achieved for those values. Furthermore, the average errors could best be interpreted in multiplicative terms, where on average, the prediction is different from the true value by a factor of 2.52. This further indicated that the model performs reasonably well on low values of online participation, which make up the bulk of the values.

Secondly, this thesis aimed to answer what the most important features are when predicting online participation in the context of public broadcasting, through determining which features contribute the most to the performance of the model. Through the permutation feature importance method, the most predictive features were identified. The features that showed to be most important are average historical social media engagement indicators, average historical views, duration of the episode, the channel, and the categories they belong to. It has to be noted that neither the direction of the effect nor an indication of the effect size can be identified using this method, only whether it changes the error of the prediction.

Thirdly, this thesis aimed to answer whether useful features of online participation could be extracted using topic modeling on the subtitles of programs in the context of public broadcasting by examining whether there a significant improvement in the R^2 and *RMSE* when including the features extracted from topic modeling. Including the topic distributions from topic modeling as features in the selection of models and tuning them accordingly resulted in an increase of 1.4% in terms of R^2 and a decrease of 5.1% in terms of *RMSE* based on back-transformed predictions. However, when comparing the cross-validation scores the performance is slightly worse when the topic features are included, therefore questioning the validity of these results. Considering that the performance increase on the test set is small, it is likely that including the topics have very little effect on the prediction performance of the model. Furthermore, only a few of the topics showed up in the top-20 important features, further indicating that the inclusion of topics is that beneficial to the model.

Fourthly, this thesis aimed to answer whether topics can be identified that are likely to generate high online participation on social media in the context of public broadcasting by determining which topics contribute most to the prediction of online participation. Several important and unimportant topics were identified using the permutation feature importance method. Tour de France, Weather and Business were found to be the most important interpretable topics and Sustainability, Politics and Christianity were found to be the most unimportant. However, overall, the differences between topics in terms of feature importance are small. Thus, the extent to which an episode is about a certain topic likely does not make large changes in terms of online participation, but there are slight differences.

Theoretical Implications

Earlier in this thesis, the problems of public broadcasting in the current social media landscape have been discussed. In the literature, the need to redefine public broadcasting has been put forward and some scholars propose that social media platforms should be embraced, for the benefit of better interactions with citizens (Murdock, 2005; Moe, 2008; Coleman, 2004; Bardoel & Lowe, 2007). This thesis argues that, as it remains unclear how the success of this should be measured, online participation could serve as an alternative, social media metric that aligns with public broadcasting's traditional aims and draws support from more general public research, where it has been identified to be a driver of engaging citizens (Mergel, 2013), and audience research, where the new forms of inquiry of social media have been found to create useful insights for broadcasting (Van Es et al., 2016). Furthermore, in the context of public broadcasting, this thesis identified that the goal of online participation is to engage citizens to participate with public broadcasting content that is meant to entertain, inform, and educate.

The results of this thesis indicated that both historical social media features and historical average views were the most important predictors of online participation. The results indicate that episodes with higher historical social media engagement and historical average views are on average, associated with higher online participation. Hence, as these were found to be the main drivers of online participation, it seems that commenting online on the social media channels of the NPO is largely a result of the general popularity of the show rather than specific characteristics of the content of the show; if a show has gained popularity online and has established a large audience of viewers, this is what drives online participation more than other characteristics of episodes. This can likely also be partly attributed to the algorithmic ranking of content (Baym, 2013). If the content of channels of NPO shows on Facebook gain popularity, their content will be more likely to be recommended to users, thereby increasing online participation. Furthermore, as the Facebook channels of NPO gain more fans, which is a form of subscribing to their content, more users will automatically find that content on their timeline. This emphasizes the importance of an effective social media campaign, as this could potentially accelerate this process. Nevertheless, the workings of the Facebook algorithm cannot fully explain how general popularity drives commenting behavior, as there need to be initial surges in popularity that it can amplify.

Earlier in this thesis, the differences between the topics generated from topic modeling and the categories have been discussed. The topics generated from topic modeling can identify more specific themes of the episodes and due to the topic distributions that were used as features, episodes can have multiple topics, as opposed to the category features. Although the topics and categories did not show to be the main drivers of online participation, there are still some slight differences in feature importance. The feature importances indicate that the episode categories Other Non-Fiction, Opinion, News & Actualities, Knowledge are the most important episode category features, and the five most important topics are the Tour De France, Weather, Business, Dutch Soccer National and Media Consumption. As the correlations are positive, a positive effect can be expected of these features on online participation, meaning that episodes of these categories and about these topics are on average, associated with higher online participation.

These most important categories and topics suggest a blend of entertaining, informative and educative content. For example, the category Knowledge implies educative content, the topic Tour De France implies more entertaining content, and the category News & Actualities implies more informative content. Thus, the categories and topics of the episodes that generate the most online participation do not show a pattern in terms of a specific type of content. Rather,

the episodes with the highest online participation are associated with a blend of different types of content. As public broadcasting's aim is to engage its viewers to engage with their content that is meant to entertain, inform, and educate, the results suggest that online participation is an appropriate metric of broadcasting success as all three types of content are associated with the highest online participation compared to other features. If we would have found that, for example, only entertaining content was important for predicting online participation compared to other types of content, it could be questioned as to whether online participation should be considered an alternative metric of broadcasting success, due to the misalignment with public broadcasting's traditional goals.

Because the topic features and categories are similar in that they both describe the type of content of a show, it could be that once the topic features were included in the model, prediction performance did not increase significantly because the categories already largely covered the information that is included in the topic features. This is also indicated by the feature importance results, where some topics replaced categories in the top-20 most important features, while most other important features did not change that much. This could therefore explain why no significant increase in performance was found and therefore why the topic features did not prove very beneficial to model performance. Nonetheless, differences between topics could be identified and therefore these topics help increase our understanding of online participation as an alternative metric of broadcasting success.

Thus, this thesis contributes to the literature on redefining public broadcasting (e.g. Van Dijck & Poell, 2015; Moe, 2008) through characterizing online participation as an alternative success metric for broadcasting success. The results suggest that online participation is mainly driven by the general popularity of the show, accelerated by algorithmic ranking. Furthermore, as all three types of public broadcasting content are associated with the highest online participation of episodes, the results indicate that, with respect to public broadcasting's traditional aims, online participation can be considered a useful measure of public broadcasting success. Finally, although the categories of episodes may already contain most information that topics generated from topic modeling contain in terms of predictive performance, due to the advantages of topics over categories, they can create some insights in terms of which topics generate more online participation.

Furthermore, this thesis extends the literature on the applications of topic modeling (e.g. Ramamonjisoa, 2014; Hu & Ester, 2013, Zhao et al., 2011; Gao et al., 2018), by introducing topic modeling as a method to extract predictive features in the prediction of online participation in public broadcasting. An LDA topic model was trained on the subtitles of

episodes to see if they can improve the prediction of online participation. As the performance increase on the test set is small, and the cross-validation results indicate a different effect, it is likely that including the topics has very little effect on the prediction performance of the model. Hence, the results of this thesis indicate that the application of topic modeling did not prove to be a way to extract predictive features for the prediction of online participation in public broadcasting.

Additionally, this thesis examined the concept of online participation, a concept that appears mostly in social science literature (e.g. Malinen, 2015; Livingstone et al., 2005), in the context of public broadcasting. In this context, episode-related features were considered in the analysis of online participation, thereby introducing a new application of online participation, and a new set of predictors specific to this context, to the work that examines online participation and its drivers. In addition, this thesis contributes to this literature by measuring to which extent online participation can be predicted and which predictors of online participation are important in this context.

Practical Implications

For public broadcasting organizations, the results of this thesis are relevant in several ways. Firstly, by proposing online participation as an alternative metric of broadcasting success, public broadcasting organizations may use online participation as a new key performance indicator to measure the success of their media content on the basis of the arguments presented in this thesis. Furthermore, as this thesis has demonstrated that online participation can be predicted with reasonable error for episodes with low online participation, considering that the error will be significantly higher for episodes with extremely high values for online participation, public broadcasting organizations may use the similar models used in this thesis to predict the expected online participation of new episodes and adjust episodes or other factors when desired.

In addition, public broadcasting organizations may consider the most important features that were determined in this thesis, to improve their predictions of online participation. The results indicated that both historical social media features and historical average views are the main drivers of online participation and therefore should be included in models that aim to predict online participation. When predicting the online participation of new shows, online broadcasting organizations should consider that the general popularity of the show weighs higher than any other factor. As a result, public broadcasting organizations may benefit from investing in their social media campaign to increase online participation and accelerate it through algorithmic ranking.

Finally, public broadcasting may use topic modeling to identify topics that are more likely to generate more online participation than others. However, they should consider that topic features are not likely to significantly improve prediction performance, especially when episode categories are already included. Therefore, public broadcasting should expect only slight differences in online participation between episodes of different topics.

Limitations and Future Research

One limitation of this thesis is that by taking the count of comments as a metric for public broadcasting success, comments that differ in interpretation are treated as homogeneous comments when composing them into this metric (Gerlitz, 2017). Therefore, important differences between the individual comments are neglected, such as what the comment consisted of, whether it was about the content of the episode or not, whether it was positive about the episode or negative, who the writer of the comment is, what the writers' intentions were with the comment, how many replies the comment received and the characteristics of these replies. These aspects may differ significantly per comment, and when composing all comments and treating them equally by counting them, they are being considered similar, while they may not. These differences between comments could potentially explain why there was still significant error in the model. For example, by making a distinction between comments that are about the content of the episode and not, Facebook users that did not watch the episode but replied for other reasons can be filtered out, which makes it more likely that the characteristics of an episode such as the topic can predict online participation. Furthermore, identifying properties of the comments like the average reply rate on comments could also turn out to be an additional valuable metric, as this implies that discussion is going on between users, instead of users just stating an opinion. In any case, collecting more information about the individual comments and then making decisions for the metric on the basis of these characteristics is likely to yield a more reflective, and potentially more predictable, metric of online participation. Hence, future research should investigate online participation as an alternative metric of public broadcasting success by collecting more information about the individual comments to see whether this can be used to develop a better metric.

Another limitation of this thesis is that online participation as a metric is non–representative of the public broadcasting audience and its engagement as a whole (Baym, 2013). In other words, the viewers that comment on social media are not representative of all viewers. This happens in two ways. Firstly, there are multiple social media platforms, and the sample of Facebook comments is an unrepresentative sample of the social media platforms as

a whole because there are differences between the users on Facebook and the users on other platforms. For example, there may be more younger users on Instagram while there may be more older users on Facebook. Secondly, when considering all social media channels, the sample is still unrepresentative because of the distinct characteristics of viewers that engage online. For example, the viewers who engage online are likely to be more outspoken than the viewers who do not engage online. Either way, this limits the results of this paper because if changes to episodes were made on the basis of online participation, changes are being made on the basis of an unrepresentative sample. For example, Business showed up as a relatively important topic, but it cannot be assumed that this is the case for all viewers and on all social media platforms, because the sample is unrepresentative of other platforms and viewers as a whole. Therefore, if changes to episodes are made on the basis of the important features in the model, it should be considered a change for a specific audience, the users on Facebook, and not viewers as a whole. Although finding a representative sample for all viewers using social media metrics is difficult, future research could adopt a similar approach as this thesis but use social media metrics of different social media platforms, so the differences between these platforms can be assessed.

Ethical Considerations

As explained in the Data section, the data used for answering the research questions do not track the individual behavior of viewers and no personally identifiable information is being collected, so the risks of misuse of personal information of users are mitigated. Furthermore, the interpretation of the outputs of the algorithms used are not likely to affect the rights or privacy of individuals in any way. As discussed in the previous section, the users that comment on Facebook are not representative for all viewers as a whole. This may have some implications in the context of ethics as well. For example, if certain groups of people are more likely to express themselves online, and this is related to social factors or ethnicity, then if certain adjustments to episodes were made on the basis of online participation, some groups of individuals may be neglected because their preferences might not be measured by the algorithms used. However, this is not a very severe or likely consequence.

Sources

Baym, N. K. (2013). Data not seen: The uses and shortcomings of social media metrics. *First Monday*.

Benington, J. (2011). From private choice to public value. *Public value: Theory and practice*, 31-51.

Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly*, *29*(1), 30-40.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.

Chen, J., Xu, H., & Whinston, A. B. (2011). Moderated online communities and quality of user-generated content. *Journal of management information systems*, *28*(2), 237-268.

Chun, S. A., & Luna-Reyes, L. F. (2012). Social media in government. *Gov. Inf. Q.*, *29*(4), 441-445.

Coleman, S. (2004). From service to commons: Re-inventing a space for public communication. *From public service broadcasting to public service communications*, 88-99.

Coulter, K. S., & Roggeveen, A. (2012). "Like it or not": Consumer responses to word-of-mouth communication in on-line social networks. *Management Research Review*.

Criado, J. I., Sandoval-Almazan, R., & Gil-Garcia, J. R. (2013). Government innovation through social media. *Government Information Quarterly 30*(4), 319–326.

Ekman, J., & Amnå, E. (2012). Political participation and civic engagement: Towards a new typology. *Human affairs*, *22*(3), 283-300.

Enli, G. S. (2008). Redefining public service broadcasting: Multi-platform participation. *Convergence*, *14*(1), 105-120.

Gao, W., Li, P., & Darwish, K. (2018). Joint topic modeling for event summarization across news and social media streams. In *Social Media Content Analysis: Natural Language Processing and Beyond* (pp. 321-346).

Gerlitz, C. (2017). 17. Data Point Critique. In *The Datafied Society* (pp. 241-244). Amsterdam University Press.

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of computer-mediated communication*, *17*(3), 319-336.

Harrigan, P., Evers, U., Miles, M., & Daly, T. (2017). Customer engagement with tourism social media brands. *Tourism management*, *59*, 597-609.

Holtz-Bacha, C., & Norris, P. (2000). *"To Entertain, Inform and Educate": Still the Role of Public Television in the 1990s?*. Joan Shorenstein Center on the Press, Politics and Public Policy.

Hu, B., & Ester, M. (2013, October). Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 25-32).

Hudson, S., Huang, L., Roth, M. S., & Madden, T. J. (2016). The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors. *International Journal of Research in Marketing*, *33*(1), 27-41.

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society*, *18*(3), 502-520.

Livingstone, S., Bober, M., & Helsper, E. J. (2005). Active participation or just more information? Young people's take-up of opportunities to act and interact on the Internet. *Information, Community & Society*, *8*(3), 287-314.

Lowe, G. F., & Bardoel, J. (2008). *From public service broadcasting to public service media: RIPE@ 2007*. Nordicom, University of Gothenburg.

Malinen, S. (2015). Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in human behavior*, *46*, 228-238.

May, A. L. (2010). Who tube? How YouTube's news and politics space is going mainstream. *The International Journal of Press/Politics*, *15*(4), 499-511.

Mergel, I. (2013). A framework for interpreting social media interactions in the public sector. *Government information quarterly*, *30*(4), 327-334.

Moe, H. (2008). Dissemination and dialogue in the public sphere: a case for public service media online. *Media, Culture & Society*, *30*(3), 319-336.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/

Murdock, G. (2005). Building the digital commons. *Cultural dilemmas in public service broadcasting*, 213-31.

Power, D. J., & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of decision systems, 20*(3), 249-261.

Ramamonjisoa, D. (2014, March). Topic modeling on users's comments. In *2014 Third ICT International Student Project Conference (ICT-ISPC)* (pp. 177-180). IEEE.

Rice, L. L., Moffett, K. W., & Madupalli, R. (2013). Campaign-related social networking and the political participation of college students. *Social Science Computer Review*, *31*(3), 257-279.

Sukumaran, A., Vezich, S., McHugh, M., & Nass, C. (2011, May). Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3401-3410).

Syvertsen, T. (2004). Citizens, audiences, customers and players: A conceptual discussion of the relationship between broadcasters and their publics. *European Journal of Cultural Studies, 7*(3), 363-380.

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and communication, 1*(1), 2-14.

Van Dijck, J., & Poell, T. (2015). Making public television social? Public service broadcasting and the challenges of social media. *Television & new media, 16*(2), 148-164.

van Es, K., van Geenen, D., & Boeschoten, T. (2015). Re-imagining Television Audience Research: Tracing Viewing Patterns on Twitter. *M/C Journal*, *18*(6).

Wang, X., & Bryer, T. A. (2013). Assessing the costs of public participation: A case study of two online participation mechanisms. *The American Review of Public Administration*, *43*(2), 179-199.

Warren, A. M., Sulaiman, A., & Jaafar, N. I. (2014). Social media effects on fostering online civic engagement and building citizen trust and trust in institutions. *Government Information Quarterly*, *31*(2), 291-301.

Zhang, W., Johnson, T. J., Seltzer, T., & Bichard, S. L. (2010). The revolution will be networked: The influence of social networking sites on political attitudes and behavior. *Social Science Computer Review*, *28*(1), 75-92.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.