Universiteit Utrecht

MSc. Applied Data Sciences

# Master Thesis

Prediction of diversity in NPO shows

Michael Wekking

## Abstract

The value of diversity, in terms of representation of people, has recently come to the forefront for public broadcasters, including the Dutch NPO. The NPO measures diversity through a questionnaire, which asks people to what extent they see or hear people from different population groups in an episode. This thesis aims to predict this ‘diversity score’ using TF, TF-IDF and LDA, to gain insight into the predictive capacity of words and topics for diversity in media content. Both words and topics are found that predict this measure of diversity: the diversity score can be predicted with explained variances between 8% and 49.7%, depending on the dataset.

## Chapter 1. Introduction

### *1.1 Motivation and context*

In the last few years, public broadcasters in Europe have increasingly focused on the representation of different people in their content, which they often describe as ‘diversity’. The European Broadcasting Union (2021), which is an alliance of 112 broadcasting organizations in 56 European countries, has published a report about diversity and the public service media, in which they state that research on the representation of minorities points to underrepresentation of certain groups and the prevalence of stereotypes. According to the EBU, ‘As research on the so-called Contact Hypothesis shows that increased contact with minority groups helps improve attitudes towards them, PSM have an essential responsibility in broadcasting more inclusive content’ (European Broadcasting Union, 2021, p. 2).

The Dutch public broadcaster, NPO, has also placed more emphasis on diversity in recent years; in their 2016-2020 policy plan, they included diversity in their list of "public values" for the first time. From then on, NPO would evaluate their content on the extent to which it contributes to the representation of different population groups (NPO, 2015). However, Kartosen-Wong (2020) argues that the plans to become more diverse and inclusive have not yet succeeded: he argues that Dutch people of non-Western origin are still insufficiently represented in NPO's shows, and that staff diversity is inadequate.

Diversity is often measured through a count of the representation of different population groups (EBU 2021; BBC 2022; de Swert et al. 2020; Daalmans & ter Horst 2017), usually measuring diversity in terms of at least one, but often more, of the following: ethnic background, gender, age, sexual orientation, and handicap. The NPO measures diversity in a different way: it measures the extent to which shows achieve public value (among which is the value of diversity), by asking the public in a questionnaire. This paper aims to be an exploratory research as to whether diversity in media content can be predicted from text elements; it seeks to predict the 'diversity score' that the public gives to a show, by means of a regression using the words and topics of the subtitles of the NPO shows as features. For the NPO, a good working model that predicts diversity from text could evaluate new shows that are pitched on how diverse they are.

### *1.2 Literature Review*

#### *1.2.1 NPO as a public broadcaster*

The NPO (Nederlandse Publieke Omroep) is a state-funded broadcasting organization in the Netherlands. It consists of the NPO as a governing body and a wide range of broadcasters, whose role is to represent the public (NPO, n.d.). The NPO coordinates the programming of all platforms (e.g., television and radio stations, social media) and the broadcasters are responsible for the content of the programs on NPO's platforms. This system has its roots in the early 20th century, when there were four broadcasters, each representing a "pillar" of Dutch society: Catholic, protestant, liberal and socialist (Bardoel, 2003).

As it is government funded, NPO is obligated by law to adhere to certain 'public values'. These are established by the government in the Mediawet, which, among other things, states that the content of the NPO needs to balanced, pluriform, varied, of high quality, accessible to everyone, and independent of politics and commerce (Mediawet, 2008). NPO has integrated these obligations in their policy plan for the period of 2022-2026. This policy plan includes seven public values on which the NPO evaluates its content: reliability, diversity, variety, independence, pluralism, personal relevance, and societal relevance (NPO, 2020).

The focus on public values was already present when public broadcasters had a broadcasting monopoly: according to Bardoel and Brants (2004, p. 167) "their right to exist was built on obligations to society in which information, quality, cultural enrichment and independence from state and commercialism were the central ingredients". With the introduction of commercial broadcasters in the 1980s and 1990s, the claim to public funding of public broadcasting became more controversial (Papathanassopoulos & Negrine, 2011, p. 25). The research of Bardoel and Brants (2004, p. 181), shows that public broadcasters have learned to legitimize their right to exist by formulating their mission "less in terms of organizational design, and more in terms of the program content they offer and the role they play in society”. This focus on public values rather than profit is what distinguishes NPO from commercial broadcasters. NPO's most recent policy plan states that "not profit, but value is leading" (NPO, 2020). Whereas commercial broadcasters focus more on providing entertainment to the public in order to make a profit, public broadcasters are focused on public missions such as educating the public and representing different groups in society.

### *1.2.2 The public value of diversity*

In the policy plan for 2022-2026, the NPO's main ambition is to provide "qualitative, multicolored and valuable content" (NPO, 2020). With the word "multicolored", the NPO alludes to the total of three of their previously mentioned public values: diversity, variety and pluralism. Diversity refers to people; NPO defines the public value of diversity as "our content represents Dutch society in gender, age, education, geographical distribution, ethnicity and disability" (NPO, 2020, p. 15). Variety refers to a diversity in subjects, and pluralism to a reflection of the diversity in opinions and ideologies in society. This thesis will focus on diversity in terms of representation of people; when diversity is mentioned in this paper, it refers to diversity in the context of representation.

The media have an important role in dealing with diversity. According to Fürsich (2010), ‘representation in the media … creates reality and normalizes specific world views or ideologies’. Thus, how certain population groups are represented in the media can have an important effect on how these groups are perceived and treated. Cultural differences exist in Western multicultural societies, and the question is not whether the media should respond to them, but how the media deals with them

(Horsti, Hulten & Titley, 2014). Stevenson (2003, p. 47) writes the following about the consequences of whether or not a person feels represented: ‘Our integrity as human beings … is dependent upon processes of cultural domination (being represented as inferior), non-recognition (being excluded from the dominant imagery of one’s culture) and disrespect (being continually portrayed in a negative or stereotypical way)’. Therefore, the issue of representation in the media is of great importance.

Still, the focus of the NPO on diversity in terms of representation is new: diversity was not mentioned as a public value by the NPO until 2016. The main concern with regards to diversity of the NPO in the years before was on pluralism: representing different ideologies in Dutch society (Engelbert & Awad, 2014). This focus on pluralism instead of diversity in the early 2010s is noticeable in the policy plan of the NPO for the period of 2010 to 2016. It is briefly mentioned that the NPO aims to offer interesting content for all population groups, thereby differentiating population groups in terms of demographics (gender and ethnicity), but also in terms of lifestyle and media consumption. Still, it is clear that the focus of the NPO was not on diversity. Contrary to ‘pluralism’, ‘diversity’ was not yet mentioned as one of the public values.

However, a shift in focus can be seen in the policy plan of 2016 to 2020. This document reads that “it is essential to connect different population groups … it is essential that diversity becomes an integral part of our programming process” (NPO, 2015, p. 43). From 2016 onwards, diversity is included as one of the public values of the NPO. The focus on diversity is further emphasized in the policy plan of 2022 to 2026, in which it is stated that the NPO will report on the opinion of the public about the diversity, pluralism, and variety of their content every year (NPO, 2020).

### *1.2.3 Measuring diversity*

The increased focus of public broadcasters on improving the diversity in their content raises the question: how can diversity be measured? Studies that measure diversity often link back to two different ways of interpreting representation: how often are people from certain population groups represented in the media and how are they represented?

Many studies have been conducted on the former; for example, the Belgian public broadcaster VRT has assigned researchers the task to count the diversity in terms of ethnic background, gender, age, and handicap every year (de Swert et al., 2020). Similarly, as part of the 50:50 project, the BBC counts the occurrence of women, people with a minority ethnic background, and people with a handicap in their content every year, aiming to increase their occurrence in BBC content (BBC, 2022).

Daalmans and ter Horst (2017) did a similar analysis for Dutch prime-time television: they analyzed the representation of gender, age, ethnicity, and sexual orientation on Dutch television by counting how many people from each of these groups were on prime-time television, and how large the proportion was of the different groups in each genre. They found that there was an underrepresentation of women, elderly people, and sexual minorities on prime-time television.

Furthermore, numerous studies have been done on how minorities are represented in the media, for example measuring how often they are portrayed as experts, in a positive or negative role, or whether they are stereotyped (Panis et al., 2019; CSA, 2021; Creative Diversity Network, 2021).

The NPO has a different way of measuring diversity: this is done through the Publieke Waarde Monitor, which is a questionnaire that is conducted every day on a panel of 9000 people in the Netherlands. This panel reports which shows they watched and to what extent these shows satisfied the public values of the NPO (NPO, 2018). The questionnaire includes people being asked to what extent a show is diverse, which is phrased as follows: ‘You hear/see different population groups’. Respondents indicate to what extent they agree with this statement; based on this, the diversity of a show is calculated (hereafter called the ‘diversity score’).

This approach to measuring diversity has two main advantages compared to the more common way of counting the representation of certain population groups. Firstly, it allows for comparing shows in terms of diversity, because this questionnaire outputs a score between 0 and 100 for how diverse a show is. When counting the representation of certain population groups, it is much harder to compare between individual episodes, because diversity is measured in different dimensions (e.g., amount of people with and without minority ethnic background, male and female, heterosexual or homosexual). Through this diversity score (and the other public values scores) the NPO can compare shows on their fulfilment of public values (NPO, 2018).

Secondly, counting the representation of certain population groups can result in a narrow definition of diversity: for example, in the study of VRT (de Swert et al., 2020) representation is measured in terms of ethnic background, gender, age, and handicap. This leaves out other differences between people that are relevant for both the VRT and NPO, such as regional background, socio-economic status, and sexual orientation. The same is true for other studies on counting representation of different population groups (BBC, 2022; Daalmans & ter Horst, 2017; EBU, 2021). In this light, the fact that the question in the NPO questionnaire is open to interpretation can be an advantage: diversity is defined through what the public finds diverse and does not have this narrow fixation on certain population groups.

However, there are also disadvantages to measuring diversity in this way. Firstly, it remains unclear from this metric which population groups are represented. The goal of the NPO is to represent all Dutch people in their content, regardless of age, gender, education, geographical spread, ethnic background, or handicap. The fact that the question about diversity in the NPO questionnaire on public values is open to interpretation also constitutes a risk: it does not specify certain population groups. Thus, the results of the questionnaire may yield high diversity scores for most programs because people believe that different population groups are represented, while at the same time one pillar of diversity may be ignored (for example, there may be too few people with disabilities in NPO content).

Secondly, due to this way of measuring diversity, a show may not have a high diversity score because it does not represent multiple population groups, even though it contributes to the representation of a specific group. For example, the NPO show ‘de Roze Supporters’ may not have a high diversity score because it represents only one specific population group (LGBT football supporters). However, this population group may not have been represented yet in other NPO shows, in which case the program contributes to the overall representation of different populations in NPO content.

The topic of measuring diversity in text elements, which this thesis also intends to do, was previously addressed in a paper by Nguyen et al. (2011), who sought to predict the age of authors through linear regression with word frequencies and POS tags as features. This study yielded positive results, with models for the different datasets with explained variance ($R^2$) ranging between 28 and 55 percent. The most predictive features were both nouns that are associated with younger and older people (e.g., ‘grandchildren’, ‘daughter’, ‘mom’, ‘school’) and words used in colloquial speech such as ‘like’, ‘just’, and ‘had’.

### *1.3 Research question*

This thesis aims to investigate whether the diversity score in the Publieke Waarden Monitor can also be predicted in this way. Returning to the question in the questionnaire, this reads: ‘You hear/see different population groups’. This implies that people hear certain words in a show that they associate with the presence of different population groups. This thesis hypothesizes that there are textual features, in the form of words and topics, that can predict whether the public thinks a show is diverse. Firstly, these may be words and topics that denote minority groups such as ethnic minorities (e.g. a word like ‘migrant’ or a topic about migration) or people from the LGBT community (e.g. words like ‘transgender’ or a topic about people with a homosexual orientation), because minorities may be perceived by the respondents of the questionnaire as a clear ‘different population group’. If the appearance of someone from a minority group in a show is combined with the appearance of someone who is not from that minority group, this can lead to a high diversity score. Furthermore, there might be words or topics that indicate the appearance of many people in one episode, for example words or topics about a city or big events. These text features could be predictive of the diversity score as well.

The research question is: *To what extent can the diversity score that the public gives to a show be predicted by the word frequencies and topics in the subtitles, and what are the most predictive features?* The research is performed by transforming the subtitles into respectively word counts (also called TF), TF-IDF and LDA topics, and predicting the diversity score with the words and topics as features using Lasso regression. I will look at the most predictive features to make sense of how the model has created the prediction.

This thesis adds to the body of literature on diversity in public service media; it is an exploratory study of the relationship between text elements and diversity in media content and whether the former can predict the latter. Furthermore, it is of practical relevance for the NPO because it can be a first step towards creating a model that recognizes the diversity of shows by its subtitles, which can be useful for evaluating scenarios of possible new shows on how ‘diverse’ they are and suggesting modifications for shows to improve their diversity.

## Chapter 2. Data

### *2.1 Subtitles and diversity score*

To conduct this analysis, the subtitles of shows of the NPO and their corresponding diversity scores will be used. In the datasets of the NPO, this data is available from 1 January 2018. In order to provide the machine learning models with enough data to predict the diversity score, all data was used from 1 January 2018 until 20 May 2022, when the data was retrieved.

With regards to the subtitles, most shows are pre-recorded; the subtitles are made by the subtitles department of the NPO before the show is broadcasted (NPO, n.d.). However, there are also shows that are broadcasted live, such as the NOS Journaal, sport broadcasts, and talk shows like Op1 and Goedemorgen Nederland. These are more prone to errors; the person who makes the subtitles does not have as much time to achieve optimal accuracy. Nevertheless, these subtitles are often changed after the episode has aired to attain more accuracy for viewers of On Demand or repeat broadcasts on television. Another limitation is that the subtitles are usually in Dutch, even when English is spoken in the show. Therefore, word meaning can sometimes get lost in translation.

The diversity score is obtained through a questionnaire called the Publieke Waarde Monitor. This is a questionnaire that is sent to 9000 people in the Netherlands, claimed as being representative of the whole population (NPO, 2018). The respondents report daily what shows they have seen and how much these shows achieve each of the public values. One of the statements that the respondents need to answer is ‘I hear/see different population groups’. The answers to this question form the dependent variable of the models used in this thesis. This variable has a score between 0 and 100. The possible answers to this question are ‘I completely agree’, ‘I mostly agree’, ‘I mostly don’t agree’, ‘I completely don’t agree’, ‘I don’t know’, and ‘This statement does not apply to this episode’. The first answer means a score of 100, the second a score of 66, the third a score of 33, and the rest of the answers a score of 0. The average of these scores is the ‘diversity score’ for an episode.

Only shows that have at least 30 respondents were included for this analysis. This number was chosen for multiple reasons. Firstly, the NPO itself only evaluates shows that have at least 30 respondents on the public values questionnaire. Secondly, a number of less than 30 respondents would make the variability of the diversity score too high to make a comparison between the different episodes. In the original public values data, there are many shows that have scores of 100 or 0 since these shows have been rated only once or twice. A dataset of at least 20 respondents per episode already yields unrealistically high and low diversity scores. Furthermore, having a larger threshold of 40 or 50 respondents would mean that too many episodes would be excluded.

### *2.2 Used datasets*

Besides performing the analysis on the remaining dataset as a whole, two subsets of the data were used as well. The first of these subsets is a dataset with one episode per show. This dataset was used because a regression based on the word count, TF-IDF, or topic modeling might focus too much on recognizing which show it is instead of demonstrating elements of diversity in shows. Some shows have a higher average diversity score than others (for example, the 'NOS Journaal' has an average diversity score of 79.6, whereas average score of all episodes is 75.1, as can be seen *Table 6* and *Table 1*). This can result in features with positive coefficients that identify the NOS Journaal, while they cannot be used for predicting diversity in other shows. Thus, to remove this bias in the features of the model, a subset of the data was created with one episode per show. The dataset was obtained by taking a random sample of one episode per show from the original dataset.

The last dataset consists of episodes from the 8 PM news. This dataset was chosen as a case study, to see if the models used in this thesis would be able to differentiate on diversity scores within a show. This particular show was chosen for two reasons. Firstly, the 'NOS Journaal 20u' is the show with the most records in the dataset; the 8 PM news accounts for 1561 of the 10709 records in the original dataset, and thus has by far the most data available for training a model. Secondly, the news usually covers major topics (e.g., the war in Ukraine, COVID) over the course of many episodes. The topics and words that are associated with a higher and lower diversity score can be derived by this analysis.

For each dataset that was used, I obtained the distribution and descriptive statistics of the diversity score and the word count of the records in the corpus. These are shown in *section 2.3*.

### *2.3 Descriptive statistics*

### *2.3.1 Descriptive statistics of dataset with all records*

**Figure 1**
*Diversity score distribution for dataset with all records*

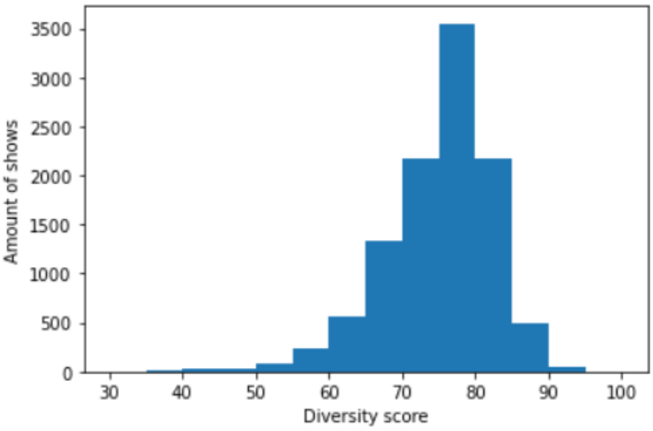

**Table 1**
*Descriptive statistics of diversity score for dataset with all episodes*

| | |
|---|---|
| Count | 10709 |
| Mean | 75.1 |
| Standard Deviation | 7.3 |
| Minimum | 34.3 |
| Q1 | 71 |
| Median | 76.6 |
| Q3 | 80 |
| Maximum | 95 |

**Figure 2**
*Word count distribution for dataset with all records*

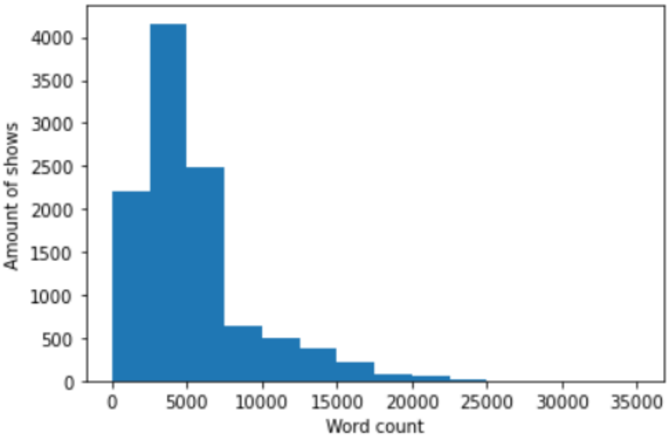

**Table 2**
*Descriptive statistics of word count for dataset with all episodes*

| | |
|---|---|
| Count | 10709 |
| Mean | 5317 |
| Standard Deviation | 3778 |
| Minimum | 0 |
| Q1 | 2868 |
| Median | 4177 |
| Q3 | 6526 |
| Maximum | 35112 |

The diversity score of this dataset somewhat resembles a normal distribution (*Figure 1*) but is slightly skewed to the left. This is not an issue for Lasso regression since the model does not assume a normal distribution of the dependent variable. The word count distribution (*Figure 2*) is skewed to the right; half of the values are between 2868 and 6526 (*Table 2*), but a significant amount of values are above 10000 as well. Looking at the shows with the most words, it can be concluded that episodes of shows usually exceed 10000 words for at least one of the following two reasons.

Firstly, an episode can be much longer than average: for example, the coverage of the Tour de France on the 28th of July 2018 lasted almost six hours, and therefore the subtitles had a word count of more than 26000. Secondly, in some of the subtitle texts in the dataset the same subtitles are repeated multiple times. For example, in the show with the most word counts (which is ‘van der Laan en Woe: Pesetas’) the subtitles are repeated three times in the text.

Unfortunately, this flaw, which occurs in a small fraction of the texts, was found too late in the process of this research: thus, the repetition of subtitles could not be removed. This can slightly worsen the performance of the model based on TF. If the model is trained on a training set in which there are no shows with repeated subtitles, the score of a show in which the subtitles are repeated three times will probably be mispredicted, since the coefficients of words that are features in this model will be added or subtracted three times.

This misprediction of TF can generally be an issue for this dataset, even if the repeated subtitles would be removed, because there is a substantial difference between the word counts of the different documents. However, this is not an issue for the model based on TF-IDF representations: L2 normalization is included in TF-IDF, which normalizes for document length, meaning that the representation of features remains the same if the subtitles are repeated. With regards to topic modeling, my assessment is that the repetition of subtitles is not a big issue, since LDA is primarily concerned with how often words co-occur in different texts.

There are also texts that contain less than 100 words. This might be problematic; it will be harder for the models to predict the diversity score because they might not contain any of the words that are used as features in the models. Therefore, texts containing less than 100 words were deleted.

The shows that occur most frequently in this dataset are interesting to observe, as these shows will have a greater impact on machine learning models. The shows that appear most often are shown in *Table 3.* The NOS Journaal is by a great amount the show that appears most often in this dataset. The Journaal of 20h and 18h together account for 2790 of the 10709 records (26 percent). Because of this imbalance in number of episodes in different shows in this dataset, I chose to include a dataset with one episode per show for my analysis as well.

**Table 3**
*Programs that appear most frequently in the dataset with all shows*

| Title | Count | Mean diversity score |
|---|---|---|
| NOS Journaal 20u | 1561 | 79.6 |
| NOS Journaal 18u | 1229 | 77.8 |
| Eenvandaag | 480 | 77.3 |
| Studio Sport Eredivisie | 287 | 70 |
| Nieuwsuur | 243 | 76.7 |
| Met het mes op tafel | 234 | 70 |
| DWDD | 230 | 75.7 |
| Op1 | 211 | 75.1 |
| De Slimste Mens | 193 | 68.4 |
| Opsporing verzocht | 174 | 84.7 |

*2.3.2 Descriptive statistics of dataset with one episode per show*

**Figure 3**

*Diversity score distribution for dataset with one episode per show*

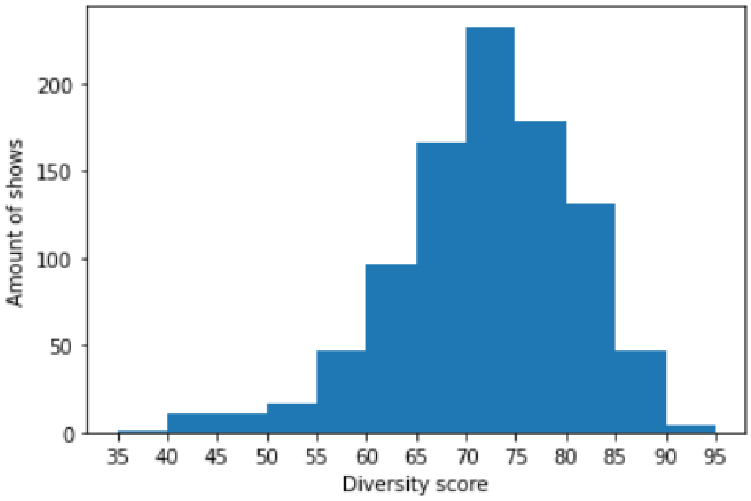

**Table 4**

*Descriptive statistics of diversity score for dataset with one episode per show*

| | |
|---|---|
| Count | 945 |
| Mean | 71.9 |
| Standard Deviation | 39.1 |
| Minimum | 34.6 |
| Q1 | 66.9 |
| Median | 72.6 |
| Q3 | 78.2 |
| Maximum | 92.8 |

**Figure 4**

*Word count distribution for dataset with one episode per show*

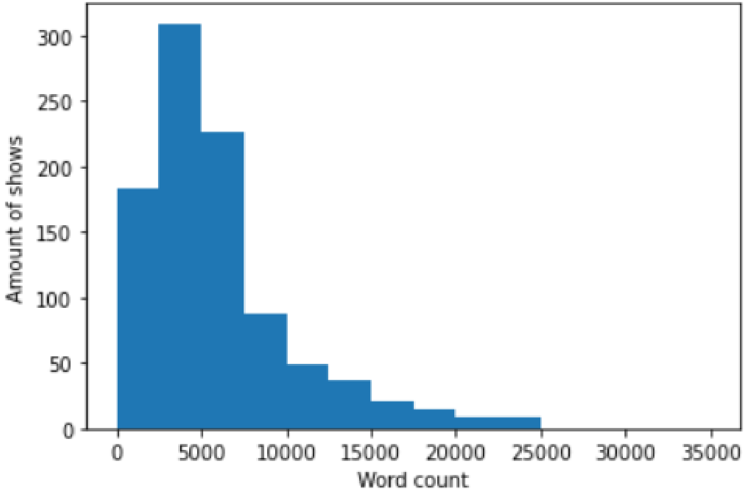

**Table 5**

*Descriptive statistics of word count for dataset with one episode per show*

| | |
|---|---|
| Count | 945 |
| Mean | 5940 |
| Standard Deviation | 4572 |
| Minimum | 76 |
| Q1 | 2801 |
| Median | 4815 |
| Q3 | 7400 |
| Maximum | 35112 |

The distribution of the diversity score in this dataset (*Figure 3*) resembles a normal distribution as well. However, *Table 4* shows that the mean diversity score is much lower (71,9 compared to 75,1) and there are far fewer values between 75 and 80 compared to the original dataset. This has to do with the fact that the shows that are rated most often (and thus appear much more often in the original dataset), like the NOS Journaal, generally have a higher diversity score than the mean, as can be seen in *Table 3*. The distribution of the word count (*Figure 4*) is right skewed in this dataset as well, and has a very large standard deviation (*Table 5*). Again, this is particularly problematic for the model based on TF transformations of the words.

### *2.3.3 Descriptive statistics of Journaal dataset*

**Figure 5**
*Diversity score distribution for Journaal dataset*

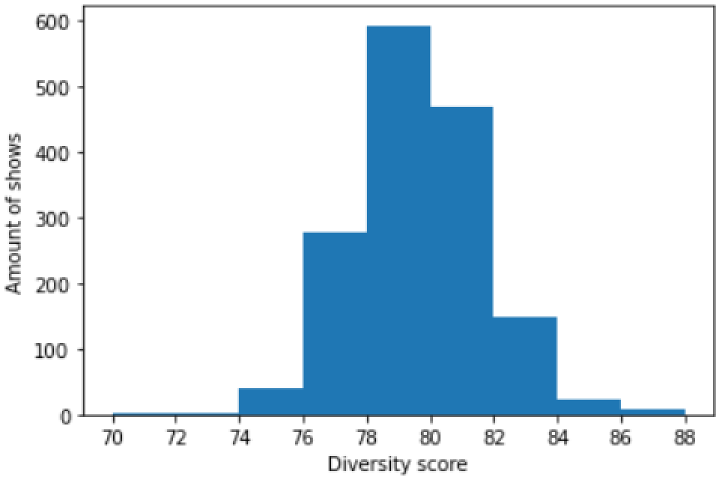

**Table 6**
*Descriptive statistics of diversity score for Journaal dataset*

| | |
|---|---|
| Count | 1561 |
| Mean | 79.6 |
| Standard Deviation | 2 |
| Minimum | 70.3 |
| Q1 | 78.2 |
| Median | 79.5 |
| Q3 | 80.9 |
| Maximum | 86.9 |

**Figure 6**
*Word count distribution for Journaal dataset*

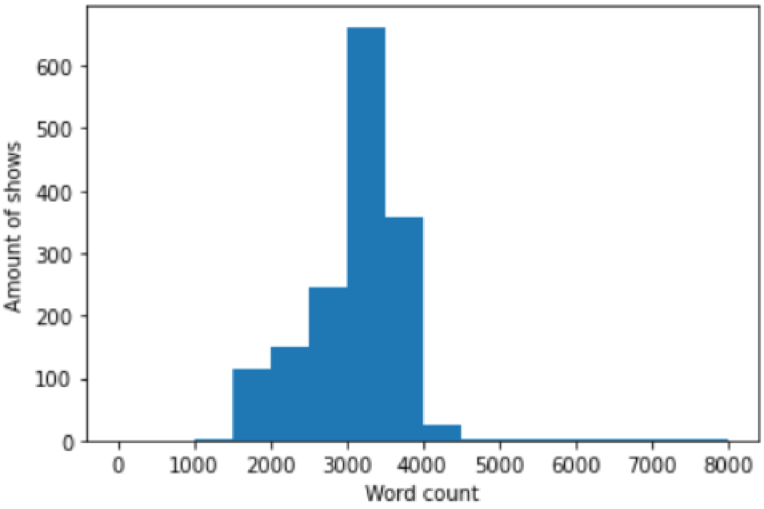

**Table 7**
*Descriptive statistics of word count for Journaal dataset*

| | |
|---|---|
| Count | 1561 |
| Mean | 3100 |
| Standard Deviation | 603 |
| Minimum | 1380 |
| Q1 | 2769 |
| Median | 3246 |
| Q3 | 3498 |
| Maximum | 7800 |

The distribution of the diversity score of the Journaal dataset, like the distributions of the other datasets, is also close to a normal distribution (*Figure 5*). However, the standard deviation of this diversity score is much lower; the values are much closer to each other (*Table 6*). This may make it more difficult for the model to differentiate between the episodes: the differences are more likely to occur due to chance (due to a few people more answering ‘I totally agree’ instead of ‘I somewhat agree’) than in the other two datasets. The values of the Journaal’s word count are also closer together, as visible in the low standard deviation compared to the other datasets (*Table 7*).

### *2.4 Episodes with highest and lowest diversity scores*

To get a better understanding of what the diversity score of the public values questionnaire is based on, I explored the highest and lowest rated episodes of the dataset with all the episodes, which can be seen in *table 8*.

**Table 8**
*Episodes with highest diversity scores*

| Title | Date | Number of respondents | Diversity score |
|---|---|---|---|
| Klassen | 4 January 2021 | 33 | 95 |
| Gevoel van de vierdaagse | 18 July 2019 | 55 | 94.1 |
| Over mijn lijk | 25 January 2018 | 51 | 93.7 |
| Floortje naar het einde van de wereld | 24 July 2018 | 30 | 93.5 |
| Over mijn lijk | 4 January 2018 | 51 | 93.1 |
| Eindelijk thuis | 18 February 2019 | 39 | 92.9 |
| Andere tijden special | 6 December 2019 | 31 | 92.9 |
| Wat een stel | 13 January 2022 | 31 | 92.8 |
| Spoorloos | 22 January 2018 | 60 | 92.3 |
| Spoorloos | 8 January 2018 | 63 | 92.2 |

When looking at the episodes with the highest diversity scores, one common denominator can be found. All episodes portray different population groups in one episode, if ‘different population groups’ is defined as the NPO defines it in its policy plan: different in terms of gender, age, education, geographical distribution, ethnicity, and disability (NPO, 2021). For example, in the “Klassen” episode, children in two classrooms are shown: one classroom consists of children with no immigrant background, the other consists mainly of children with an immigrant background. Children from both classes are interviewed. In "Gevoel van de vierdaagse" several people participating in the "Vierdaagse" are shown and interviewed; they are people of different ages, genders, and regional and national backgrounds. Finally, in 'Wat een stel' several couples are followed for a year, ranging from a couple with an immigrant background to gay couples and a couple with Down's syndrome. This exploration of the highest diversity scores indicates that the diversity score on the public values questionnaire measures what it is supposed to measure, which is whether people from multiple population groups can be seen or heard in an episode.

**Table 9**
*Episodes with lowest diversity scores*

| Title | Date | Number of respondents | Diversity score |
|---|---|---|---|
| Journaal extra | 14 January 2022 | 311 | 34.3 |
| Paulien Cornelisse – Om mij motiverende redenen | 12 April 2020 | 30 | 34.6 |
| Nationaal aftelmoment | 31 December 2020 | 68 | 35.6 |
| Journaal extra | 25 January 2022 | 289 | 36.3 |
| Journaal extra | 15 February 2022 | 158 | 36.6 |
| Prof. Mr. Pieter is 80 | 30 April 2019 | 38 | 37.6 |
| Theo Maassen - Vankwaadtoterger | 14 June 2020 | 56 | 38 |
| Journaal extra | 12 November 2021 | 207 | 38.2 |
| Journaal extra | 14 December 2020 | 346 | 39.6 |
| Media Inside | 1 May 2022 | 37 | 39.9 |

Most episodes among the ones with the lowest diversity scores are episodes of ‘Journaal extra’. These are all broadcasts of press conferences of the Dutch government in which they talk about measures related to COVID. The people speaking at these press conferences are always two of the following four men: Mark Rutte, Hugo de Jonge, Ernst Kuijpers, and Jaap van Dissel. These are all men who are over 40 years old, which is also a trend that can be found in the rest of the shows with the lowest diversity scores, apart from Paulien Cornelisse’s comedy show (a white woman 46 years of age). From this it can be concluded that the respondents of the questionnaire consider shows with only one or more white men to be the low point of the current definition of diversity.

### *2.5 Ethical and legal considerations*

The data that the NPO collects is confidential; to get access to the data, a non-disclosure agreement (NDA) needed to be signed. In this agreement, it is stated that data cannot be shared with third parties and may only be used in the context of this thesis.

The data used for this research does not concern any personal data; the only data that was retrieved from people was through the questionnaire of the NPO, but this data is aggregated and personal information is not visible in the results of the survey. Thus, issues related to misuse of personal information do not come up in this thesis.

The biggest ethical considerations in this thesis lie in the effects that a model that automates the measurement of diversity in a show might have in practice. These will be discussed in *section 5.5*.

## Chapter 3. Methods

### *3.1 Translation to data science question*

The research question of this paper is: *to what extent can the diversity score that the public gives to a show be predicted by the word frequencies and topics of its subtitles, and what are the most predictive features*? This research question implies a regression problem since the dependent variable (the diversity score) is a numerical variable. In order to answer the question, two methods are used: a regression based on TF and TF-IDF, and a regression based on topics that are derived from a topic modelling approach called LDA. The question was converted into the following data questions:

1. What is the $R^2$ for the regression models with as features respectively the words and the topics?
2. What are the words and topics that are most predictive for the diversity score?

For both the regression based on TF and TF-IDF, and the regression based on topic modeling, I will explain how I performed the analysis and why I chose these methods.

### *3.2 Text Regression*

A collection of texts, like the subtitles used in this thesis, is an unstructured dataset. To use this data in a machine learning model, the data needs to be converted into a structured dataset by creating features that function as independent variables for the model. Two of the methods that are often used for creating these features are Term Frequency (TF), also called word count, and Term Frequency Inverse Document Frequency (TF-IDF).

A Term Frequency model encodes texts into a large vector, with the words that occur in all the texts as features. The vector indicates how often each word occurs in the given text. TF-IDF is a combination of TF with IDF (Inverse Document Frequency). IDF measures how common a word is in the full set of documents. It is calculated by dividing the total number of documents by the number of documents in which the word appears and taking the logarithm of this number. The advantages of TF-IDF over IDF are twofold. Firstly, it reduces the effect of frequently-used words in the model. Secondly, TF-IDF allows for the application of normalization techniques such as L2 normalization, which is used in this analysis. L2 normalization is a normalization technique that modifies the values of the dataset such that in each vector the sum of the squares equals 1. This is especially important in the context of this thesis because texts with vastly differing word counts are used (see *section 2.3*). In this case, a TF representation can make two documents appear different because they have different lengths, but the length does not contribute much to understanding the meaning of the document. Normalization, as used in TF-IDF, removes the effect of the document length.

TF and TF-IDF are commonly used for text classification tasks. These techniques are easy to use and valuable for prediction tasks such as the task at hand in this thesis: they can extract the most descriptive terms in a document and using these, explain the variance in the dependent variable to a certain extent. Text classification based on TF and TF-IDF can be used for tasks such as sentiment analysis (predicting whether a text expresses a positive or negative sentiment by looking at the words), or dividing texts into categories (for example, dividing the news into categories like 'politics' and 'sports' based on keywords). It was also used successfully in regression tasks such as the paper by Nguyen et al. (2011), where the age of the author is predicted through word counts in the text. This thesis aims to perform a regression task as well: predicting the diversity score of programs using TF and TF-IDF.

However, there are limitations to using these techniques. For example, they do not work well in determining the meaning of a word within a sentence. The sentences "this restaurant is great" and "this restaurant is not great", while describing the opposite sentiment, are similar in TF; "restaurant" and "great" have a word count of 1 in both sentences. Furthermore, words that refer to the same thing (for example 'bike' and 'bicycle') are treated as different words in TF and TF-IDF.

The texts were preprocessed by tokenization, removing punctuation, and lowercasing all words. Words that were very uncommon in the datasets (less than 5 times in the dataset with all records, less than 3 times in the other datasets) were removed; these can be typos and otherwise do not add much to the predictive power of the model due to their infrequency. Lemmatization was not used for this model, because it causes a loss of information in the texts; it is no longer possible to separate present and past tense or to tell apart different conjugations of words. Initial model runs showed that lemmatizing features did cause a loss of information and worsened the performance of the model.

Deleting stop words may also cause a loss of information for the model: stop words can potentially say something about the diversity of a program. For example, the words ‘wij’ and ‘we’ indicate that the person speaking is part of a group. If these words are used often in a show, this can mean that there are multiple population groups in this program. To avoid losing potential valuable information for the model, the stop words were not removed initially. All models were run with and without stop words, to see which version of the model performs better.

### *3.3 Topic modeling*

Topic modeling was performed using Latent Dirichlet Allocation (LDA). LDA is a method first mentioned in Blei et al. (2003), in which they describe it to be a way of dimensionality reduction that preserves the essential statistical relationships useful for tasks such as classification. This way of dimensionality reduction can yield better results in classification tasks than using TF (Karioglu et al. 2013) or TF-IDF (Blei et al. 2003). Resnik et al. (2013) performed topic modeling and used it as an input for a regression task; they found that it produced interpretable themes that added value to predictions. Because of the satisfactory performance of LDA in classification and regression tasks, and because topics may yield new insights on the relation between text elements and the diversity score, LDA was used in this thesis.

LDA assumes that a document is composed of a distribution over topics and that each topic is a distribution over words. Because of the Dirichlet distribution, it is assumed that a document consists of more than one topic, but not many different topics. In turn, each topic consists of a distribution over words, with some words occurring a lot within that topic and others occurring little or not at all. These distributions are obtained as follows: the words are first randomly assigned to one of the k topics. Then, the document-topic and topic-word distributions are optimized by looking for a mapping in which words that occur together are classified within the same topics as much as possible. The model outputs the selected number of topics as a distribution over words, and each document as a distribution over topics. These obtained topics are used as independent variables for the machine learning model, and the topic probabilities of each document are used to predict the dependent variable.

A major criticism of LDA is that it yields unstable topics (Vayansky & Kumar 2020, Agrawal et al. 2018). The words are at first randomly assigned to one of the k topics; thus, different runs of the model will have different initializations of words to topics and will therefore have varying results. Furthermore, LDA assumes that topics are independent of each other, while correlation between topics is according to Vayansky and Kumar (2020) "a common part of many types of data, especially text … data". Vayansky and Kumar argue that "this limits the ability of this algorithm to handle big data accurately and make predictions for new documents". Similarly, LDA assumes that documents and words within documents are independent, which is usually not true.

Again, the first steps of preprocessing were tokenization and removing punctuation and capitalization in the texts. For LDA, stop words were deleted, because these words are not useful for extracting topics from the texts. If the output of LDA would contain topics with words such as 'de' and 'een', this would not be helpful in determining what subjects are talked about in the episodes. Words that occur often in the texts (more than 35 percent of the time) were also deleted; these do not contribute to a division into many distinct topics of the texts. Words that occur infrequently in the texts are also not useful for dividing the texts into topics: if words occur very infrequently, they cannot be part of a significant topic in the texts. Therefore, words that occur in less than 10 of the texts in the 'one episode per show' dataset and Journaal dataset, and words that appear less than 20 times in the 'all episodes' dataset were deleted. Finally, both single words and bigrams were extracted from the text, so that for example 'united states' will be seen as one feature.

Subsequently, LDA was run in the package gensim on the preprocessed texts and the topic distributions were extracted from each document. The package gensim was chosen over the sklearn package since the sklearn package, in an initial model run, yielded topics with words that did not seem to be related to each other. The number of topics for each dataset was chosen by qualitative analysis of different amounts of topics (10, 20, 30, 40, 50, and 60) to see which number produced the most meaningful topics.

### *3.4 Lasso Regression*

These features (respectively TF, TF-IDF, and the topic distributions) were subsequently used as inputs for a Lasso regression model. Lasso regression is a type of linear regression which includes a L1 penalty: this causes some coefficients in the model to become zero. How many coefficients become zero is determined by the parameter lambda, which was tuned in this analysis by investigating at what value of lambda the model did best in cross-validation. Lasso regression was used because the models (especially the models based on TF and TF-IDF) would otherwise have an excessive number of variables and thus would overfit on the training set. Lasso regression ensures that the variables with the best predictive power for the training set are chosen and many other variables are set to zero.

# Chapter 4. Results

The results are divided into two sections: one for the regressions based on TF and TF-IDF and the other for the regressions with topic distributions derived from topic modeling as features.

## *4.1 TF and TF-IDF*

*Table 10* shows the results of running each of the Lasso regression models (using TF and TF-IDF, with and without stop words) for the different subsets of data, documenting for each model the explained variance ($R^2$) of the dependent variable of the training set, and the $R^2$ derived by cross-validation. The best performing model in cross-validation was evaluated using a test set.

**Table 10**
*Performance of Lasso regression models based on TF and TF-IDF on different datasets*

| Dataset | Train/CV/Test $R^2$ | TF with stop words | TF without stop words | TF-IDF with stop words | TF-IDF without stop words |
|---|---|---|---|---|---|
| All records | Train | 0.542 | 0.528 | 0.679 | 0.678 |
| | CV | 0.396 | 0.397 | 0.486 | 0.486 |
| | Test | | | | 0.497 |
| One episode per show | Train | 0.39 | 0.587 | 0.668 | 0.709 |
| | CV | 0.265 | 0.249 | 0.271 | 0.275 |
| | Test | | | | 0.257 |
| Journaal | Train | 0.115 | 0.08 | 0.081 | 0.082 |
| | CV | 0.044 | 0.054 | 0.048 | 0.057 |
| | Test | | | | 0.08 |

The results from *table 10* show that the TF-based models perform slightly worse than those based on TF-IDF, which was expected due to the different text lengths. Furthermore, the exclusion of stop words usually improved the TF-IDF models. A general pattern that can be observed from this table is that the models overfit: generally, the $R^2$ of the model on the training set is at least a factor of 1.5 higher than the $R^2$ on the test set. This means that some features do not have the predictive power that the model ascribes to them on the test set.

The following tables (*table 11, 12, and 13*) show the words with the highest coefficients in the best predictive models for each of the datasets. These predictive models were all derived using TF-IDF. The TF-IDF of words in documents are multiplied with the coefficients of these words and added to the intercept to obtain the diversity score. To explain why these words have high positive and negative coefficients, I looked back at the texts in which they appear most often and looked for common trends in these texts.

**Table 11**
*Words with highest positive and negative coefficients for Lasso regression model of dataset with all episodes*

| Words with positive coefficients | Coefficient | Words with negative coefficients | Coefficient |
|---|---|---|---|
| nieuwsuur | 50.2 | themakanaal | -439.1 |
| lucia | 48.4 | rossems | -260.2 |
| vierdaagse | 43 | lubach | -114.7 |
| alfabet | 33.5 | vrouwtjes | -89.7 |
| reid | 30.8 | 2022 | -73.2 |
| praten | 30.5 | profeet | -62.4 |
| oase | 30.4 | roofdieren | -60.6 |
| meegenomen | 30 | tomorrow | -53.1 |
| armen | 28.6 | haaien | -48.4 |
| wanneer | 27.8 | gelach | -46.5 |

I divided the words in *Table 11* into 2 categories: words that are fully or mostly connected to one show, and words that occur in multiple shows:

**Category 1** (words that are fully or mostly connected to one show)**:** nieuwsuur, lucia, alfabet, reid, oase, themakanaal, rossems, lubach, tomorrow

**Category 2** (words that occur in multiple shows): vierdaagse, praten, meegenomen, armen, wanneer, vrouwtjes, 2022, profeet, roofdieren, haaien, gelach

The words that are fully or mostly connected to one show are useful predictive features in the model because there are some shows that have significantly higher (e.g., ‘NOS Journaal’, ‘Oase in de Oriënt’) or lower (e.g., ‘Hier zijn de van Rossems’, ‘Zondag met Lubach’) diversity scores. These are not the most interesting features for the purpose of predicting the diversity of a NPO show, because it only demonstrates that these specific programs have higher or lower diversity scores.

A more useful word for this purpose is ‘vierdaagse’. The episodes where the word ‘vierdaagse’ appears most are not only from the show ‘Het Gevoel van de Vierdaagse’; the word also appears in other episodes that have a high diversity score, for example in the show ‘Jinek’ and the 6 PM news. This predictive power of the word ‘vierdaagse’ on the diversity score likely has to do with the fact that the Vierdaagse of Nijmegen is an event that a diverse group of people visits: people from different areas of the country, with varying ages, men and women, etc. Another predictive word with a positive coefficient is ‘praten’: the frequency of this word is highest in fiction shows, the news, and talk shows. It signifies human interaction and thus the presence of multiple people and possibly different population groups. As for words like ‘meegenomen’, ‘armen’, and ‘wanneer’, the reasons for the predictive capacity of these words are not clear: these appear in many different shows which do not share an obvious connection.

In the negative coefficients, a common pattern can be seen in three words: ‘vrouwtjes’, ‘roofdieren’ and ‘haaien’. These words occur most frequently in documentaries about animals. The other words that have a negative coefficient are ‘2022’, ‘gelach’ and ‘profeet’. The word ‘2022’ as a negative coefficient indicates that shows broadcasted in the year 2022 had a lower diversity score on average. The word ‘gelach’ usually occurs in subtitles of shows that have a live audience; it refers to the audience laughing. These shows (among others ‘de Avondshow met Lubach’, ‘Media Inside’, and ‘Doorbakken’) generally have a lower average diversity score. This might be because these shows usually do not include many people: only the host and one or two guests. The word ‘profeet’ is also notable in the negative coefficients; a positive coefficient would be expected because of its association with muslims. However, two of the shows where the word ‘profeet’ are used most often are a comedy show by Theo Maassen, in which he ridicules Islam, and an episode of ‘Zondag met Lubach’, where Arjen Lubach talks about teachers showing cartoons of the prophet Mohammed in class. In both shows only one white man is on screen, hence the diversity score is low.

**Table 12**

*Words with highest positive and negative coefficients for Lasso regression model of dataset with one episode per show*

| Words with positive coefficients | Coefficient | Words with negative coefficients | Coefficient |
|---|---|---|---|
| wonen | 39.9 | tomorrow | -42.8 |
| lopen | 26.6 | gelach | -42.8 |
| goedemorgen | 26.2 | mos | -41 |
| mensen | 24.7 | weet | -39.5 |
| jaar | 21.1 | rivier | -39 |
| lang | 20.6 | economie | -37.9 |
| hallo | 18.9 | panda | -36.7 |
| chinees | 18.3 | bijen | -35.9 |
| dag | 18.1 | trio | -35.1 |
| hans | 17 | hond | -34.8 |

**Category 1 (Greeting people):** hallo, goedemorgen

**Category 2 (Travel or exploring places):** wonen, lopen, lang, chinees

**Category 3 (Animals and nature):** rivier, panda, bijen, hond, mos

In the words with the highest positive coefficients for the ‘one episode per show’ dataset (*table 12*) trends can be found as well. Most words with positive coefficients can be classified into Category 1 (greeting people) or category 2 (travel or exploring places). The words in category 1, ‘goedemorgen’ and ‘hallo’, are both used for greeting people. These words appear most often in shows where a host meets different people (e.g., ‘Over Mijn Lijk’, ‘Thuis op Zuid’, ‘Hello Goodbye’). It seems logical that in shows where many people are greeted, there is a higher likelihood that this show includes people from different population groups.

The word with the highest coefficient in this model is ‘wonen’. This word usually occurs in travel shows (‘Floortje naar het einde van de wereld’, ‘Erica op Reis’) or shows about a particular place in the Netherlands (‘Typisch den Dolder’, ‘Kolping, een volkswijk in renovatie’). These shows typically show multiple people or groups of people that live in that place; ‘wonen’ may therefore be a predictor of the diversity score. Another word that usually refers to this type of show is ‘lang’: this word often refers to the history of a place (e.g., ‘2500 jaar lang’). The word ‘lopen’ sometimes occurs in travel shows as well, but also in Athletics events such as the marathon of Rotterdam and the

Olympics, which both scored high on the diversity measure. The word ‘chinees’ occurs most often in shows where the presenter travels to China and meets different people there (e.g., ‘Chinese Dromen’, ‘Door het hart van China’).

Furthermore, the word ‘jaar’ often appears in the coverage of events that take place every year (e.g., ‘Canal Parade 2019’, ‘Kerst muziekgala 2021’, ‘de Nationale Dodenherdenking’). At these events, generally many different people are present; these programs may have a higher diversity score for this reason. Finally, the word ‘mensen’ frequently appears in a context where general remarks are made about a group of people (for example in ‘Minister van Gehandicaptenzaken’, ‘Canal Parade 2019’, and ‘Levenslucht: een week op de IC tijdens Corona’). These people that are talked about are usually also depicted on screen.

As for the words with negative coefficients in *table 12*, a pattern can be found in words that indicate animals and nature, which was classified as category 3 (with words like ‘rivier’, ‘panda’, ‘bijen’, and ‘hond’; and ‘mos’ in some contexts). These words indicate that the object of interest in the episode is something related to nature; it therefore seems logical that not many people will be featured in these episodes. Another interesting word is ‘economie’: this word occurs most often in the context of politics (e.g., ‘Prinsjesdag’, ‘het EenVandaag Verkiezingsdebat’, ‘EenVandaag: de Politieke Prestatie’). The diversity score that the public attributes to these shows is lower than average, indicating that the respondents of the questionnaire do not find politicians or the hosts that cover politics in TV shows diverse. Furthermore, the words ‘tomorrow’ and ‘gelach’ emerge as negative coefficients again. ‘Gelach’ was explained in the section about *table 11*, ‘tomorrow’ occurs as a negative coefficient mostly because one episode of ‘Promenade’, which has a diversity score of 43, features the word ‘tomorrow’ often, because it features a song in which this word is prominent.

**Table 13**
*Words with highest positive and negative coefficients for Lasso regression model of Journaal dataset*

| Words with positive coefficients | Coefficient | Words with negative coefficients | Coefficient |
|---|---|---|---|
| btw | 4 | corona | -8 |
| oorlog | 3.8 | vertrouwen | -3.7 |
| oekraïense | 1.8 | coronacrisis | -3.6 |
| maan | 1.6 | bibi | -3.6 |
| johnson | 1.2 | hulp | -3.3 |
| iran | 0.9 | grot | -3 |
| notre | 0.9 | winkels | -2.5 |
| oekraine | 0.9 | kabinet | -2.5 |
| suriname | 0.8 | gaan | -2.5 |
| blok | 0.8 | coronamaatregelen | -2.3 |

**Category 1 (war, Ukraine):** oorlog, oekraïense, oekraïne

**Category 2 (Other countries):** oekraïense, johnson, iran, notre, oekraine, suriname

**Category 3 (COVID)**: corona, coronacrisis, winkels, coronamaatregelen

**Category 4 (Politics)**: corona, vertrouwen, coronacrisis, winkels, cabinet, coronamaatregelen

The words with positive coefficients in *table 13* were divided in two categories that are not mutually exclusive: the war in Ukraine and other countries. It can be concluded that episodes of the news that include the war or other countries often have a higher diversity score. This can possibly be explained by the fact that the news covers multiple items: if there is an item about another country, this usually appears next to items about events in the Netherlands. In this way, there are multiple population groups present in the episode.

Common trends in the negative predictors of *table 13* are COVID and Dutch politics. The fact that words related to COVID are negative predictors in this model is striking, since COVID is a global pandemic, not limited to one population group or even to the Netherlands. One explanation for COVID being a significant negative predictor, however, is that news coverage may have been predominantly focused on government action on COVID. Furthermore, words that relate to the government (‘vertrouwen’ which is most often used in connection with the government, and ‘kabinet’) are negative predictors as well.

### *4.2 Topic modeling*

The following table shows the performance of the Lasso regression models with the topic distributions of documents as features, displaying the training and test score.

**Table 14**
*Performance of Lasso regression models based on LDA topics on different datasets*

| | Train/Test | Lasso Regression $R^2$ |
|---|---|---|
| All records | Train | 0.338 |
| | Test | 0.332 |
| One episode per show | Train | 0.266 |
| | Test | 0.19 |
| Journaal | Train | 0.098 |
| | Test | 0.077 |

The results as shown in *table 14* of the regression models based on topic distributions are somewhat similar to the regression models based on individual words: the model performs best for the dataset with all the records, and worst for the dataset with only the 8 PM news. Compared to TF-IDF based model, the performance of the regression based on topic distributions is much worse on the dataset with all records, slightly worse on the dataset with one episode per show, and similar on the Journaal dataset. Although the performance of the models is somewhat worse, the models suffer less from overfitting.

*Table 15* shows the most important features for each Lasso regression model (meaning, with the five highest positive and negative coefficients). These are easily interpretable: the predicted diversity score of a show can be calculated by multiplying the coefficient with the topic distribution of a document added to the intercept. The topics were named based on the top 10 words of each topic; these can be found in the Appendix.

**Table 15**

*Most predictive topics for regression model*

| 'All episodes' dataset | | 'One episode per show' dataset | | Journaal dataset | |
|---|---|---|---|---|---|
| Topic | Coefficient | Topic | Coefficient | Topic | Coefficient |
| Travel | 22.2 | Hospital | 8 | Miscellaneous (news from abroad?) | 1.4 |
| Olympics + ice-skating | 16.5 | Human interaction | 6.8 | War + Ukraine | 0.8 |
| Foreign countries | 14.7 | Royal family + China | 6.3 | Law | 0.7 |
| Journaal? | 11.2 | Concert | 5.4 | Miscellaneous | 0.1 |
| Police | 11.1 | Christianity | 4.9 | Education | 0.1 |
| Slimste Mens + quiz | -7 | Football | -3.4 | United States | -0.2 |
| Cycling | -7.2 | Cycling | -5.8 | Miscellaneous | -0.6 |
| Nature + animals + farmers | -8.4 | Sports + Olympics + ice-skating | -6 | Miscellaneous (COVID + international politics?) | -0.6 |
| Women + pregnancy | -9.2 | Miscellaneous | -12.7 | Dutch politics | -0.7 |
| Show with audience? | -17.5 | Quiz | -16.9 | COVID | -1.8 |

As can be seen in *Table* 15, meaningful topics can be found that predict a higher or lower diversity score. Some topics and coefficients are in line with earlier findings in the TF-IDF models. For example, the topics 'Travel' (signified by words such as 'reis', 'reizen', 'vliegtuig', 'kerk' and 'dorp') and 'Foreign places' (words like 'suriname', 'eiland', 'afrika' and 'curacao') are in line with the earlier finding in the analysis of *table 12* that shows about visiting a place, either abroad or in the Netherlands, have higher diversity scores. Earlier findings are also corroborated in the negative coefficient of the topic 'Nature + animals + farmers', and in the positive coefficient of war and Ukraine and negative coefficients of Dutch politics and COVID in the Journaal dataset. The positive coefficient of the topic 'Human interaction' is somewhat in accordance with an earlier finding in *table 12* that words for greeting people predict higher diversity scores. The topic 'Human interaction' also includes the word 'hallo' in its top 10 words, and further contains words like 'haha', 'oh', and 'gezellig'.

Furthermore, this method found new predictors of the diversity score, with the police, the Olympics and the hospital as positive predictors, and quizzes and most other sports as negative predictors. Words that relate to the police most often occur in shows like ‘Ellie op Patrouille’, ‘Noodcentrale’, ‘Opgelicht’ and ‘Opsporing Verzocht’, or in the news. The former shows revolve around the police and how they deal with perpetrators and victims of crimes; these shows receive higher diversity scores than average. The victims of crimes in these shows are diverse in terms of age, gender, geographical location, and ethnic background, and the perpetrators of crimes are often middle-aged men, either with or without a migration background. Quizzes are negative predictors in two models: in the model based on the ‘all episodes’ dataset and the one based on the ‘one episode per show’ dataset. This can be explained by the fact that shows about quizzes usually feature only a few people: the presenter and the few participants.

# Chapter 5: Discussion & Conclusion

## *5.1 Answering the research question*

The two main purposes of the current study were to investigate to what extent the diversity score that the public gives to a show can be predicted by the word frequency and topics in the subtitles, and to find out what the most predictive features are for these models.

The models based on TF-IDF performed better than those based on TF, which is in line with expectations because the TF-IDF model normalizes for text length. The TF-IDF model performs best on the dataset with all records (with an $R^2$ for the test set of 0.497), followed by the dataset with one episode per show ($R^2$ of 0.257); the model for the Journaal dataset performs worst with an $R^2$ of 0.08. It seems logical that the model performs best for the 'all records' dataset, since there are many features that merely indicate specific shows, and therefore the model is especially fit to programs that appear often in the dataset (like the news). The model for the 'one episode per show' dataset indicates more general features that explain some of the variance between shows, although these features should be looked at with some caution since they explain much more of the variance in the training set than in the test set. The Journal dataset is the most difficult to predict, which can be explained by the fact that the diversity scores are much closer together and the episodes are much more similar.

The regression models based on LDA topics perform significantly worse in the 'all records' ($R^2$ of 33.2%) and 'one episode per show' ($R^2$ of 19%) datasets, and similarly ($R^2$ of 7,7%) on the Journaal dataset. However, it can be concluded that the derived topics are useful in explaining at least some of the variance in the diversity score of the datasets; the model does find topics that correlate with the appearance of different population groups. Moreover, these models do not overfit on the training set, making the most predictive features more generalizable to the test set.

In the results section, some common trends can be found in which type of words and topics yield higher and lower diversity scores, although these were not entirely in line with what the hypothesis. It was hypothesized that words and topics related to minorities would be predictors of the diversity score, but this was not found in the results.

Consistent with the hypothesis, words and topics were found related to the appearance of many people. For example, from the analysis of *table 12* and *15* it was concluded that words and topics that relate to visiting places (either in the Netherlands or abroad) are associated with higher diversity scores. The same is true for words and topics related to human interaction and big events. As negative predictors, words and topics that related to nature and animals, shows with an audience, and quizzes were frequently found. For the case study of the news, clear positive and negative predictors were found in both the words and topics. Those with positive coefficients related to Ukraine and war in general, whereas those with negative coefficients were linked with COVID and Dutch politics.

### *5.2 Theoretical implications*

The remit of the NPO, as a public broadcaster, is based on the fulfilment of public values, such as providing balanced news of high quality and representing different population groups and opinions. The representation of different groups of people in their content is an important public value, since non-recognition of populations in the media can have a negative effect on how they are seen and treated. Representation is measured by the NPO through a question in the public values questionnaire, which asks whether people see or hear different population groups in a specific episode.

This thesis aimed to explore the link between text elements and diversity in media content: can text elements be used as predictive features for diversity? It can be concluded that in the shows of the NPO, and with their measurement of diversity, word frequencies and topics are useful predictive features for diversity. The explained variances of the models were generally in the range of the study of Nguyen et al. (2011) for predicting the age of an author, indicating that the task of predicting this diversity measure from the subtitles is a fruitful undertaking, which can be explored further by improving the model. The models seem to find logical predictive words and topics for predicting the perceived appearance of different population groups in episodes.

It is likely that these models would also perform well in similar contexts if other public broadcasters would decide to measure diversity in this way as well. However, this study cannot draw a general conclusion about the predictive link between text elements and diversity in media content: whether this is possible depends on how diversity is operationalized.

### *5.3 Practical implications*

As suggested in *section 1.3*, a model that predicts the diversity of a show for the NPO can be interesting because it can help to evaluate new shows that are pitched to the Directie Video of the NPO on their representation of different population groups, and to modify current shows so that their diversity can be improved.

For the first task of evaluating new shows, the performance of the models on the ‘one episode per show’ dataset seems to be most important, because the test set of the models for the ‘all episodes’ dataset also includes shows that were already in the training set. The best performing model for the ‘one episode per show’ dataset (namely, the model based on TF-IDF features) cannot perform the first task accurately yet, since it explains 25.7% of the variance for shows in the test set. The predictive performance of the model is a good starting point but needs to be improved before it can be used for this purpose.

As for the second task, the current models are not especially suited for identifying how shows could be modified to improve their diversity. Most of the predictive features say something about the type of show and/or the topic (e.g., a travel show, coverage of a big event, a show about nature or a quiz). It is not logical to recommend for a show with a relatively low diversity score to completely change the topic or genre of their show in order to get a higher diversity score. Shows that cover a wide range of topics might be an exception; these might choose to cover a topic that is associated with a higher diversity score. The predictive feature in the model that is most feasible for shows to change is human interaction (indicated in the models by words like ‘hallo’, ‘goedemorgen’, ‘hallo’, ‘leuke’, and ‘gezellig’): the degree to which a show exhibits human interaction can be changed.

### *5.4 Limitations and further research*

The models used in this thesis are a starting point for a model that predicts the diversity score. There are several ways in which the current models could be improved: firstly, the data should be cleaned to solve the problem of repeated subtitles in a text. Secondly, the features from TF-IDF and topic modeling could be combined to achieve a more accurate prediction. Thirdly, more features could be added such as the genre of the show. The results of this study have shown that the genre of the show (for example, a travel show or a quiz) can be an important predictor of the diversity score; thus, including all genres might add towards a better prediction. Finally, instead of Lasso regression another way of feature selection combined with another type of regression (for example, a state-of-the-art regression technique such as gradient boosting) could yield better predictions.

Another important limitation lies in the measurement of diversity that the NPO uses. As discussed in *section 1.2.3*, there are disadvantages to measuring diversity in this way. This measurement shows whether people think that many different population groups are represented, but it does not show which population groups are represented. Furthermore, shows that add significantly to the representation of a specific group of people can have a low score in this measure of diversity. Thus, to know specifically which population groups are represented in which episodes, the measurement of counting the occurrence of people from specific population groups should be used.

Further research could investigate whether this measurement could be automated as well; word elements such as 'gay' might for example indicate the inclusion of people from the LGBT community. However, it might be hard to fully automate this from text elements: a differentiation must be made between other people talking about a specific population group and people from this population group being represented on screen. The fact that the word 'profeet' was used most often in shows in which only one white man appeared (which was concluded in the analysis of *table 11*) indicates that this is not an easy task.

### *5.5 Ethical implications*

A model that predicts the diversity score may be used to predict the diversity of new shows that are pitched. However, the pitfall of using such a model for this purpose, is that the model is based on one operationalization of diversity: namely, the extent to which the public thinks that different population groups are represented in episode. Thus, the model will miss information that also matters about diversity, such as which population groups are represented in this episode and how they are represented. Therefore, such a model should not be used as a definitive measurement of diversity.

## Literature

Agrawal, A., Fu, W. & Menzies, T. (2018). What is wrong with Topic Modeling? (and How to Fix it Using Search-based Software Engineering). *Inform Software Tech, 98*, pp. 74-88.

Bardoel, J. (2003). Back to the public? Assessing public broadcasting in the Netherlands. *Javnost-The Public, 10(3),* pp. 81-95.

Bardoel, J. & Brants, K. (2003). From Ritual to Reality: Public Broadcasters and Social Responsibility in the Netherlands. In Lowe, G.F. & Hujanen, T. (Eds.), *Broadcasting and Convergence: New Articulations of the Public Service Remit* (pp. 167-187). Goteborg: Nordicom.

BBC (2020, June 22). *BBC commits £100m of its content spend on diverse productions and talent*. Retrieved from https://www.bbc.co.uk/mediacentre/latestnews/2020/creative-diversity-commitment

BBC (2022). *50:50 The Equality Project Impact Report 2022*. Retrieved from https://www.bbc.co.uk/5050/documents/5050-impact-report-2022.pdf

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,* pp. 993-1022.

CSA (2021, June). *Baromètre de la représentation de la société française*. Retrieved from https://www.csa.fr/Informer/Collections-du-CSA/Observatoire-de-la-diversite/Barometre-de-la-representation-de-la-societe-francaise-resultats-de-la-vague-2020

Creative Diversity Network (2021, January 27). *Diamond: The Fourth Cut*. Retrieved from https://creativediversitynetwork.com/wp-content/uploads/2021/01/CDN-Diamond4-JANUARY-27-FINAL.pdf

Daalmans, S. & ter Horst, C. (2017). Diversity reflected? Analyzing the representation of gender, age, ethnicity, and sexual orientation on Dutch prime time television. *Communications, 42(2),* 253-268.

Engelbert, J. & Awad, I. (2014). Securitizing Cultural Diversity: Dutch public broadcasting in post-multicultural and de-pillarized times. *Global Media and Communication, 10(3).*

European Broadcasting Union (2021). *Diversity and Public Service Media*. Retrieved from https://www.ebu.ch/publications/research/membersonly/report/diversity-and-public-service-media

Fürsich, E. (2010). Media and the representation of Others. *International Social Science Journal, 61(199)*, pp. 113-130.

Horsti, K, Hultén, G. & Titley, G. (2014). *National Conversations: Public Service Media and Cultural Diversity in Europe.* Intellect Books.

Isar, Y.R. (2006). Cultural Diversity. *Theory, Culture & Society, 23.*

Kartosen-Wong, R. (2020, July 18). 'Diversiteit is bij de NPO eerder wensdenken en pr-praat dan werkelijkheid'. *Het Parool.*

Mediawet 2008. (2008). Retrieved from https://wetten.overheid.nl/BWBR0025028/2022-03-02

Nguyen, D., Smith, N.A., Rosé, C.P. (2011). Author Age Prediction from Text using Linear Regression. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities,* pp. 115-123

NPO (2009). *Concessiebeleidsplan 2010-2016: Verbinden, Verrijken, Verrassen.* Retrieved from https://over.npo.nl/organisatie/openbare-documenten/concessiebeleidsplan

NPO (2015, June). *Concessiebeleidsplan 2016-2020: Het publiek voorop.* Retrieved from https://over.npo.nl/organisatie/openbare-documenten/concessiebeleidsplan

NPO (2018). *Beleidslijn Publieke Waarden*. Retrieved from https://over.npo.nl

NPO (2020, October). *Concessiebeleidsplan 2022-2026: Van waarde voor iedereen*. Retrieved from https://over.npo.nl/organisatie/openbare-documenten/concessiebeleidsplan

NPO (n.d.) *Het publieke bestel*. Retrieved from https://over.npo.nl/organisatie/bestuur-en-organisatie/het-publieke-bestel

NPO (n.d.). *T888 Ondertiteling*. https://over.npo.nl/voor-publiek/toegankelijkheid/tt888-ondertiteling#Media

Panis, K., Paulussen, S. & Dhoest, A. (2019). Managing Super-Diversity on Television: The Representation of Ethnic Minorities in Flemish Non-Fiction Programmes. *Media and Communication, 7(1)*

Papathanassopoulos, S. & Negrine, M. (2011). *European Media.* Cambridge, UK: Polity.

Sarioglu, E., Yadav, K. & Choi, A.H. (2013). Topic Modeling Based Classification of Clinical Reports. *Proceedings of the ACL Student Research Workshop*, pp. 67-73.

Stevenson, N. (2003). *Cultural Citizenship, Cosmopolitan Questions.* Maidenhead: Open University Press.

De Swert, K., Kuypers, I. & Walgrave, S. (2021). *Monitor Diversiteit 2020: Een kwantitatieve studie naar de zichtbaarheid van diversiteit op het scherm in Vlaanderen.* Retrieved from https://vrt.be

Vayansky, I. & Kumar, S.A.P. (2020). A review of topic modeling methods. *Information Systems* (2020).

# Appendix

## Appendix 1

*Words and their weights for most predictive LDA topics for ‘all episodes’ dataset*

| Words with weights | Topic Coefficient |
|---|---|
| 0.089*"reis" + 0.063*"kerk" + 0.053*"dorp" + 0.046*"trein" + 0.036*"reizen" + 0.025*"bergen" + 0.017*"kilometer" + 0.017*"europa" + 0.016*"vliegtuig" + 0.012*"spanjaarden" | 22.2 |
| 0.082*"meter" + 0.067*"goud" + 0.050*"schaatsen" + 0.042*"medaille" + 0.041*"ijs" + 0.035*"kees" + 0.027*"finale" + 0.025*"vrouwen" + 0.025*"canada" + 0.023*"medailles" | 20.5 |
| 0.023*"suriname" + 0.016*"eiland" + 0.014*"honden" + 0.013*"afrika" + 0.010*"curacao" + 0.010*"hond" + 0.008*"surinaamse" + 0.008*"antwoord" + 0.007*"zuid_afrika" + 0.006*"categorie" | 14.7 |
| 0.006*"politie" + 0.003*"wind" + 0.003*"trump" + 0.003*"president" + 0.003*"achterlopen_goedenavond" + 0.003*"zon" + 0.002*"regen" + 0.002*"vannacht" + 0.002*"graden" + 0.002*"amerikaanse" | 11.2 |
| 0.014*"politie" + 0.007*"zaak" + 0.005*"mevrouw" + 0.005*"slachtoffer" + 0.005*"meneer" + 0.005*"beelden" + 0.004*"informatie" + 0.004*"onderzoek" + 0.004*"daders" + 0.003*"telefoon" | 11.1 |
| 0.009*"seconden" + 0.008*"ehm" + 0.006*"film" + 0.006*"ronde" + 0.005*"maarten" + 0.005*"meneer" + 0.005*"antwoorden" + 0.004*"mevrouw" + 0.003*"genoemd" + 0.003*"bel" | -7 |
| 0.014*"rijden" + 0.009*"rijdt" + 0.009*"kilometer" + 0.007*"gereden" + 0.007*"finale" + 0.006*"winnen" + 0.005*"reed" + 0.005*"poel" + 0.005*"seconden" + 0.005*"meter" | -7.2 |
| '0.014*"dieren" + 0.012*"boeren" + 0.009*"natuur" + 0.007*"eten" + 0.005*"boer" + 0.005*"grond" + 0.005*"vogels" + 0.005*"aarde" + 0.004*"zee" + 0.004*"gebied" | -8.4 |
| '0.121*"vrouwen" + 0.026*"baby" + 0.024*"maria" + 0.017*"bart" + 0.015*"kind" + 0.014*"sophie" + 0.014*"zwanger" + 0.011*"bevalling" + 0.011*"seks" + 0.009*"kindje" | -9.2 |
| '0.016*"gelach" + 0.010*"applaus" + 0.007*"programma" + 0.003*"tv" + 0.003*"hoekschop" + 0.002*"rkc" + 0.002*"woord" + 0.002*"tafel" + 0.002*"publiek" + 0.002*"trump" | -17.5 |

**Appendix 2**

*Words and their weights for most predictive LDA topics for ‘one episode per show’ dataset*

| Words with weights | Topic Coefficient |
|---|---|
| 0.007*"contact" + 0.007*"ziekenhuis" + 0.006*"kanker" + 0.006*"patienten" + 0.006*"patient" + 0.005*"gesprek" + 0.005*"dood" + 0.004*"ziekte" + 0.004*"onderzoek" + 0.004*"zorg" | 8 |
| 0.004*"eten" + 0.004*"haha" + 0.003*"oh" + 0.003*"hallo" + 0.003*"gelach" + 0.003*"hahaha" + 0.003*"ehm" + 0.003*"gezellig" + 0.002*"heerlijk" + 0.002*"bed" | 6.8 |
| 0.017*"koning" + 0.016*"maastricht" + 0.016*"stad" + 0.011*"koningsdag" + 0.009*"burgemeester" + 0.008*"limburg" + 0.007*"china" + 0.006*"afrika" + 0.006*"maas" + 0.006*"feest" | 6.3 |
| 0.009*"zingen" + 0.008*"applaus_gejuich" + 0.006*"liedje" + 0.005*"liefde" + 0.005*"lied" + 0.004*"podium" + 0.004*"the" + 0.004*"hart" + 0.004*"zingt" + 0.004*"vrijheid" | 5.4 |
| 0.029*"jezus" + 0.011*"god" + 0.008*"liefde" + 0.007*"kruis" + 0.006*"vrienden" + 0.006*"maria" + 0.006*"taart" + 0.005*"hart" + 0.005*"licht" + 0.005*"hemel" | 4.9 |
| 0.029*"bal" + 0.007*"overtreding" + 0.006*"oranje" + 0.006*"spelers" + 0.006*"scheidsrechter" + 0.005*"doelpunt" + 0.005*"engeland" + 0.005*"hoekschop" + 0.005*"duitsland" + 0.004*"memphis" | -3.4 |
| 0.015*"rijden" + 0.011*"wout_aert" + 0.011*"koers" + 0.010*"rijdt" + 0.010*"poel" + 0.009*"ploeg" + 0.009*"kop" + 0.008*"kilometer" + 0.008*"peloton" + 0.007*"mathieu_poel" | -5.8 |
| 0.011*"rijden" + 0.008*"kilometer" + 0.008*"sneller" + 0.008*"rijdt" + 0.007*"schaatsen" + 0.007*"goud" + 0.006*"gereden" + 0.006*"rit" + 0.006*"olympische_spelen" + 0.006*"seizoen” | -6 |
| 0.022*"gelach" + 0.004*"youp" + 0.004*"tv" + 0.003*"haha" + 0.003*"programma" + 0.003*"sinterklaas" + 0.003*"amsterdam" + 0.003*"peter" + 0.003*"meneer" + 0.003*"zaal" | -12.7 |
| 0.063*"gelach" + 0.026*"punten" + 0.020*"team" + 0.014*"antwoord" + 0.009*"ronde" + 0.007*"punt" + 0.007*"fout" + 0.007*"goede_antwoord" + 0.007*"volgende_vraag" + 0.005*"finale" | -16.9 |

## Appendix 3

*Words and their weights for most predictive LDA topics for Journaal dataset*

| Words with weights | Topic Coefficient |
|---|---|
| 0.003*"israel" + 0.003*"china" + 0.003*"trump" + 0.002*"vs" + 0.002*"parijs" + 0.002*"demonstranten" + 0.002*"amerika" + 0.002*"brand" + 0.002*"kerk" + 0.002*"militairen" | 1.4 |
| 0.016*"rusland" + 0.012*"oekraine" + 0.010*"russische" + 0.008*"russen" + 0.007*"oorlog" + 0.006*"poetin" + 0.004*"kiev" + 0.004*"oekraiense" + 0.003*"gas" + 0.003*"navo" | 0.8 |
| 0.005*"advocaat" + 0.005*"moord" + 0.004*"rechtbank" + 0.004*"verdachte" + 0.003*"openbaar_ministerie" + 0.003*"vrouwen" + 0.003*"verdachten" + 0.003*"justitie" + 0.003*"advocaten" + 0.002*"jos" | 0.7 |
| 0.002*"vrouwen" + 0.002*"groningen" + 0.002*"plannen" + 0.002*"huizen" + 0.002*"gas" + 0.002*"kosten" + 0.002*"rijden" + 0.002*"wet" + 0.002*"elektrische_auto" + 0.002*"schiphol" | 0.1 |
| 0.004*"school" + 0.003*"onderwijs" + 0.003*"leerlingen" + 0.003*"vrouwen" + 0.003*"jongeren" + 0.002*"mannen" + 0.002*"geweld" + 0.002*"scholen" + 0.001*"studenten" + 0.001*"actie" | 0.1 |
| 0.014*"trump" + 0.008*"biden" + 0.006*"joe_biden" + 0.005*"republikeinen" + 0.005*"president_trump" + 0.005*"democraten" + 0.005*"vaccin" + 0.005*"verkiezingen" + 0.004*"biomassa" + 0.004*"donald_trump" | -0.2 |
| 0.002*"china" + 0.002*"jongeren" + 0.001*"vrouwen" + 0.001*"spelen" + 0.001*"frankrijk" + 0.001*"schade" + 0.001*"wedstrijd" + 0.001*"muziek" + 0.001*"burgemeester" + 0.001*"zomer" | -0.6 |
| 0.004*"ggd" + 0.003*"virus" + 0.003*"israel" + 0.002*"turkije" + 0.002*"grens" + 0.002*"migranten" + 0.002*"app" + 0.002*"testen" + 0.002*"besmettingen" + 0.002*"wet" | -0.6 |
| 0.011*"rutte" + 0.010*"cda" + 0.007*"belastingdienst" + 0.005*"debat" + 0.004*"vvd" + 0.004*"omtzigt" + 0.003*"formatie" + 0.003*"vertrouwen" + 0.003*"politieke" + 0.003*"koning" | -0.7 |
| 0.006*"ziekenhuizen" + 0.006*"virus" + 0.005*"patienten" + 0.004*"rivm" + 0.004*"coronacrisis" + 0.003*"lockdown" + 0.003*"coronavirus" + 0.003*"testen" + 0.003*"besmettingen" + 0.003*"horeca" | -1.8 |