Universiteit Leiden

# Master Computer Science

[Tracing Political Positioning]


| | |
|---|---|
| Name: | [Chris Congleton] |
| Student ID: | [s2577240] |
| Date: | [18/7/2022] |
| Specialisation: | [Data Science] |
| 1st supervisor: | [Peter van der Putten] |
| 2nd supervisor: | [Suzan Verberne] |


Master's Thesis in Computer Science


Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

## Abstract

Newspapers write for a particular readership and from a certain ideological or political perspective. This thesis applies computational methods to newspaper data to analyse to which extent the ideological positioning of newspapers is reflected in their writing. Political bias is illustrated in terms of coverage bias and agenda setting. Furthermore, the use of generative models (GPT-2) is explored for this purpose. These analyses showed several indications of political tendencies: disproportionate coverage of certain politicians and parties, limited overlap of political discourse, classifiable article source and divergence of generated text thematically and in terms of sentiment. Therefore, reading a newspaper requires a critical attitude which considers the intricate political tendencies of the source.

# 1 Introduction

Newspapers typically write for a particular audience, and from a certain ideological or political perspective. For opinion articles this is not necessarily a problem as long as authors and media are transparent about their position [5], yet ideological or political bias is an issue for analysis or news reporting articles [11]. Framing the debate and setting the political agenda offers this media considerable influence according to how sceptical the reader consumes the content. The media has been visualised on political bias and news value scales to this end[1].

Specific newspapers shape their readers' view through how and to what extent they select, present and discuss political issues as a subset of the collective political discourse. Unlike modern social media where each user publishes on personal account, a newspaper is formed by the collection of articles from different writers and tied together by the editor. This editorial coherence shapes the ideology or political perspective of the newspaper[11]. By selecting what news newspapers collectively cover and how they write in terms of sentiment and theme, the scope of political discourse is determined which is the basis on which parties distinguish themselves and the public casts their vote. Inversely, politicians or parties might shape their messages into a format that make it more likely to be included.

Recent research in Germany [6], Denmark [9] and Korea [13] has quantified bias in seemingly politically neutral articles by means of modern computational techniques. With regard to selection bias work by Susanszky et al.[26] measures the extent to which demonstrations in Hungary are under reported in pro-government media outlets. Their analysis is based on a dataset containing 329 articles. Furthermore, there is research on the bias of vanilla GPT-2 in relation to occupation-gender ratio [16] as well as political bias [19]. Following up on GPT-2 with 1.5B parameters trained on 40GB of text and published in 2019, a larger model GPT-3 was published in 2020 with 175B parameters and trained of a filtered dataset of 570GB [21][4].

There is a gap between these works of specific bias analysis of a subject, media outlet or source and bias on a high level generative model like GPT-2 based on an enormous set of textual data. Therefore this thesis aims to bridge this gap by tracing political positioning of newspapers based on a large collections of their articles. Furthermore, in this thesis we aim to retain the bias in the language model and analyse it in contrast to studies reducing bias in the model.

Work by De Vries et al. [29] recycling the originally English GPT-2 model has made a Dutch version available through Huggingface. This version is partly trained on old newspaper articles for 2007. Therefore, this thesis extends on this work by fine-tuning the model on more up to date Dutch articles from 2021.

Analysing and visualising political bias, scope and coherence in newspapers can uncover and unpack political and ideological orientation. Understanding these underlying mechanisms facilitates safeguarding readers by showing the true colour of sources where this is obfuscated. Therefore, this thesis seeks to answer the following key research question and sub-questions:

- RQ: To what extent is the ideological position of newspapers reflected in their writing?
- SQ1: To what extent is coverage bias measurable in newspapers?
- SQ2: To what extent do newspaper share a set of topics?

[1] Visualisation of the position of media on a political bias and news value scale by Vanessa Otero https://adfontesmedia.com/static-mbc

- SQ3: To what extent do specific newspapers cover the shared political topics?
- SQ4: To what extent can newspapers be identified given a political article?
- SQ5: To what extent does text generation based on specific newspapers diverge in topic or sentiment?

The research questions are answered by applying computational techniques to a collection of 96,840 Dutch newspaper articles collected for the purpose of this thesis. To answer the first sub question the political bias in newspapers is quantified. The second sub question provides a high level, illustrative view of the scope of the collective political discourse present in newspapers. The third sub-question aims to uncover the specific scope of political discourse unique to newspapers compared to the shared topical perspective among newspapers. The fourth sub question looks to illustrate the coherence of articles in newspapers. Finally, for the fifth sub question we fine-tune natural language generation models on specific newspapers, to analyse the divergence in text generation on a sentiment and thematic basis.

The contribution of this thesis consists of introducing the application of computational methods to newspaper data for quantitative political research and analysis. Specifically, analysing selection bias by means of topic modelling and word embeddings, analysing identifiability of article source by means of classification models and analysing thematic and sentiment divergence by analysing the output of Dutch source specific fine-tuned GPT-2 models. Applying generative language models is unique in this context of political bias. Some research that approaches this topic is an attempt to deep fake politicians on twitter by Ressmeyer et al. and a domain specific BERT model for the 2020 election for example [22][15]. Furthermore, the size of the Dutch dataset used for these analyses is among the larger of those used in related work.

The remainder of this thesis is organized as follows. Section 2 reviews the background and related work. We then present methods, (section 3). Furthermore, an overview of the collected data is given, (section 4). This is followed by a description of the experiments and results, (section 5). Followed by a discussion (section 6) and conclusion (section 8).

# 2 Background and Related Work

In this section a background context on the study of political ideology in media is constructed by discussing bias, political ideology spaces, source classification of textual data and text generation.

## 2.1 Bias

In the context of this thesis, bias is the action of supporting or opposing a particular person or party in an unfair way by allowing opinion to influence judgment. It can manifest itself in various ways as illustrated in research by Eberl et al.[8] where political bias in media is divided into three types: visibility, tonality and agenda bias. Visibility bias is defined as the effect of a party or politician receiving a relatively undue amount of coverage. Tonality bias describes the sentiment, positive or negative, of articles towards a party or politician. Agenda bias concerns the alignment of topics or issues covered by the news and a party or politician's agenda topics. Quantifying visibility and tonality bias is a good step towards answering the first sub question of this research. A lot of work has been done to investigate visibility and tonality bias. For example, the work of Dallmann et al. [6] covering political bias in online newspaper articles uses occurrence metrics and sentiment analysis. Enevoldsen et al. [9] specifically use sentiment analysis to study tonality bias.

## 2.2 Dimensionality of political discourse

Various research studies have also been conducted on the modelling of the political ideology space. Traditionally, this space is orientated on the one-dimensional spectrum from left to right, even though this contrasts the complexity and multifaceted reality of public policy. For example, the convergence of the extremes known as the Horseshoe model, challenges this linear view by discussing a convergence of extreme right and left [27]. Similarly, the representation of politics in newspapers is not limited to a single-dimensional scale. Modern dimensionality reduction techniques have been applied to find the essential dimensions needed to distinguish party politics [1][18] using surveyed political stance data. A similar approach could be applied to newspaper data to analyse the scope of political discourse which in turn can be used to answer to what extent shared political topics are discernible [24].

Quantifying agenda bias approaches the third sub question as it covers the newspaper specific topical shape in contrast to the general political discourse. Research on agenda bias in news media is not found, and thus leaves a void to be filled by this research thesis.

## 2.3 Source identification

In order to approach the extent of coherence in writing between political articles from the same newspaper, reverse analysis is applied by developing a system for the task of articles' source classification. Research work on this topic has been performed in the context of author identification of natural language [23] using a support vector machine and deep learning based approaches. Furthermore, author identification of code using word embeddings, tf-idf and convolutional neural nets shows very accurate results [2]. Another angle of discriminating articles is whether the content has a commercial or editorial purpose. The work of Kats et al.[14] shows that it is possible to differentiate between the two with an accuracy of 90%. On the basis of these research papers, it is expected to be possible to develop an appropriate system to discriminate between articles' source and analyse on what basis they are distinguished.

## 2.4 Text generation

Modern natural language generation has many applications where new texts are generated from existing ones, text-to-text generation. However, the generation of text from other forms of data or a language model can also be required [10]. In this thesis, a transformer-based language model is utilised, specifically GPT-2 [21]. This model can be applied to many tasks for example question and answering, translation and summarisation. In this thesis, we use it as auto-regressive model, generating natural language by predicting the next word in a sequence following up a prompt. The architecture of GPT-2 closely follows the setup as described in Radford et al.[20] which is based on the Transformer model [28]. In this thesis we aim to retain and analyse the bias of a source using a generative AI model. To our knowledge no research work has been done using modern language models from this angle.

| Party | Party Leader | Seats | Orientation |
|---|---|---|---|
| *VVD* | Rutte | 34 | right, conservative |
| *D66* | Kaag | 24 | center-right, progressive |
| PVV | Wilders | 17 | far right, conservative populist |
| *CDA* | Hoekstra | 15 | center, confessional |
| PvdA | Ploumen | 9 | center-left, progressive |
| SP | Marijnissen | 9 | left, conservative |
| GL | Klaver | 8 | left, progressive |
| FVD | Baudet | 8 | far right, conservative populist |
| PvdD | Ouwehand | 6 | center-left, animal rights special interest |
| *CU* | Segers | 5 | left, confessional |
| SGP | van der Staaij | 3 | right, confessional |
| Volt | Dassen | 3 | centrist, progressive European |
| JA21 | Eerdmans | 3 | right, conservative |
| BBB | van der Plas | 1 | right, farmers special interest |
| BIJ1 | Simons | 1 | far left, progressive, inclusive |
| 50PLUS | den Haan | 1 | right, seniors special interest |

Table 1: Tweede Kamer (house of representatives): parties (government coalition in italics), party leaders, seats as of March 17th 2021, approximate orientation

# 3 Methods

In this section the methods for the experimentation are outlined by describing how Coverage Bias, General and Source Specific Discourse, News Source Discrimination and Article Generation are researched in this thesis.

## 3.1 Coverage Bias

Coverage bias (SQ1) is studied by evaluating various metrics based on the aggregation of party or politician mentions. The selection of parties and politicians of which mention frequency is counted is based on the elected parties in the Tweede Kamer in 2021 and their respective party leaders. The parties, party leaders and number of seats in the 2021 Tweede Kamer are given in Table 1. The parties that are in bold: VVD, D66, CDA and CU form the current government coalition.

### 3.1.1 Log Normalised Mention Frequency

The first metric is calculated on the complete data set. First of, the occurrence of each term (party or politician) in each article is counted. The resulting counts are aggregated by summation over each source. Furthermore, these values are normalised by dividing over the total sum of term mentions, all politicians and parties, in a specific source. This takes care of the discrepancy in number of articles per source. Finally, the logarithm of these values is taken to make the results interpretable as initially the normalised counts of lesser prominent parties or politicians are dwarfed by the greater. The formula is given in equation 1. The Log Normalised Mention Frequency is denoted with $f_{ln}$. S is the set of sources in the complete data set D. An article is denoted as a and the term count in an article is denoted by $t_a$. The T is used to denote all political terms, politician or party names.

$$f_{ln} = log(\frac{\sum_{a \in S} t_a}{\sum_{a \in S} T_a}) \tag{1}$$

### 3.1.2 Relative Normalised Mention Frequency

The second metric is computed over the complete dataset as well. First of, the occurrence of each term (party or politician) in each article is counted. The resulting counts are aggregated by summation over each source. Furthermore, these values are normalised by dividing over the total sum of term mentions in a specific source. Thereafter, in order to establish the relative term mention frequency, the average mention frequency of a term over all sources is computed and this value is used to normalise the counts per source. There are four sources thus adding the $\frac{1}{4}$ fraction to the equation.

$$f_{rn} = \frac{\sum_{a \in S} t_a}{\frac{1}{4} \sum_{a \in D} t_a} \tag{2}$$

The Relative Normalised Mention Frequency is denoted with $f_{rn}$. S is the set of sources in the complete data set D. An article is denoted as a and the term count in an article is denoted by $t_a$.

## 3.2 General and Source Specific Political Discourse

To investigate to what extent shared political topics are discernible (SQ2), as well as source specific topics (SQ3), Latent Dirichlet Allocation (LDA)[3], a modern topic modelling technique is used.

### 3.2.1 Topic Modelling

Latent Dirichlet Allocation(LDA) is a generative probabilistic model of a corpus. The assumption is made that a probabilistic distribution over latent topics can be used to represent each document. The distribution of topics in all documents has a shared Dirichlet prior [12].

To prepare the text data specifically for the topic modeling punctuation and special characters are removed and the text is lowercased. An NLTK stop word list is used to remove non-significant words. This list is extended manually to remove remaining html tags.

In addition to the preprocessing of the text, for this analysis, a subset of the total article collection is used to construct the LDA model. As the purpose of this research is to distinguish political topics, political articles are selected. An article is deemed political if it contains one of the parties or party leaders names. This selection reduces the corpus size from 96.840 articles to 15.498.

The LDA model is implemented using the Gensim package and analysed using the pyLDAvis package. An example of this representation is given in figure 1. On the left the Intertopic Distance Map is depicted and on the right the top-30 most relevant terms for a topic. A relevance metric of $\lambda = 0.3$ is used to balance the word probability under a topic relative to its lift [25]. Each topic is interpreted manually based on the top-30 most relevant terms for the topic.

The number of topics is setup consistent with the number of topics that provides the most distinguishable topics over the general data on a manual basis. Some experiments with various numbers of topics was performed ranging from 5-20. Here the clearest topics were present with the number of topics set to 10. This number is kept consistent for each of the specific sources in order to compare a set of the same size.

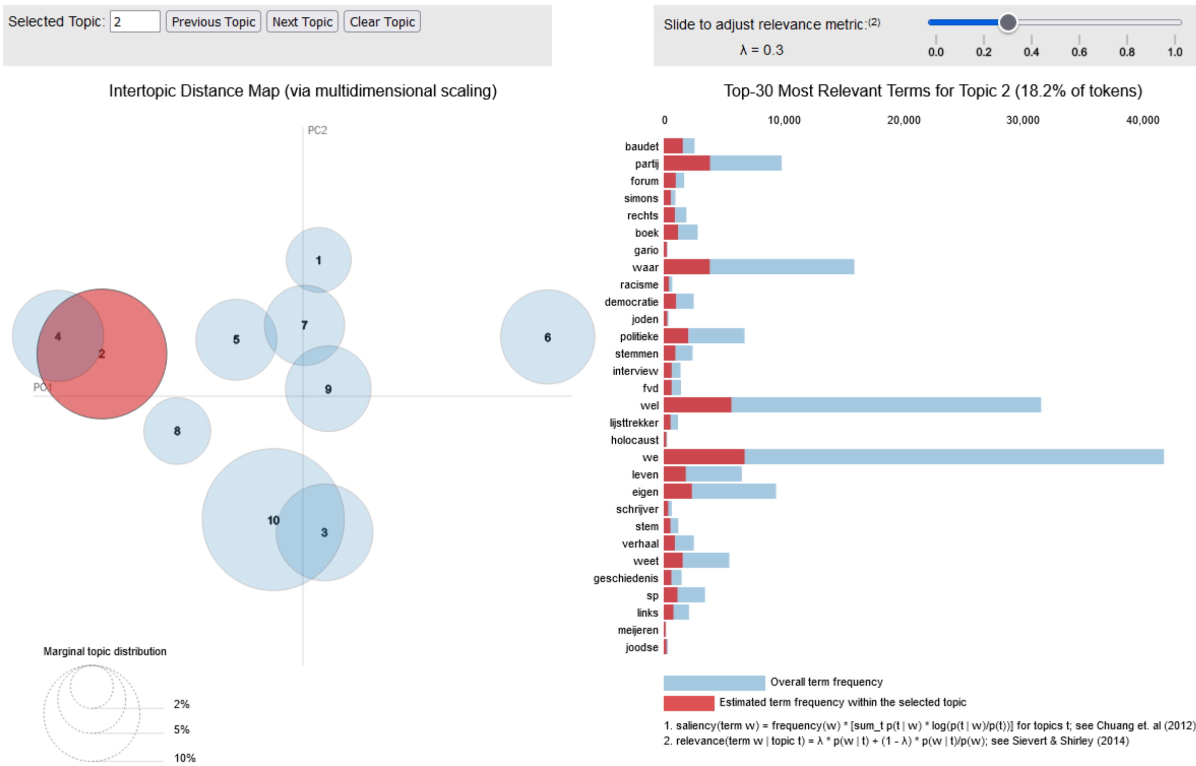


Figure 1: LDA visualisation with pyLDAvis package

### 3.2.2 Word Embeddings

Another approach to comparing the general and source-specific discourse is to represent the text data in vector space and visualise the respective embeddings of parties and politicians in a lower-dimensional space. In the embedding space words that are similar and appear in the same context have a similar vector. Visualising these vectors can therefore show what parties or politicians are discussed in a similar context. This offers a spacial projection of the parties and politicians based on how newspapers write about them as an alternative to the Horseshoe model as introduced in section 1. That model is based on the ideological position of a party or politician while this projection is based on newspaper coverage and the position in political discourse.

First, a gensim Word2Vec model is trained on the corpus of the political subset to cover general political discourse and a political subset of a newspaper to cover specific political discourse. Words representing the same party or politician are drawn together. For example, "GroenLinks" and "GL". Second, the dimensionality of the word vectors in the model is reduced with t-SNE. Third, this reduced word representation is extracted for parties and politicians and visualised.

## 3.3 Discriminating Newspapers by Article Texts

The fourth sub-question is approached by training classifier models to distinguish political articles by source and analysing the features the classifier uses to discriminate. The input of the models consists of the political subset of the data set labeled with the respective source. First, the models are tuned and compared with respect to performance. The models and classifications are then interpreted and analysed.

Determining the source identifiability is an approach to analyse the style coherence of a source. The features a model uses to distinguish sources can inform us on the major differences between sources. Furthermore, the complexity of this task says something about the depth of these difference. For example, if distinction is manageable for a simple model this would mean there is a big difference in superficial aspects of the textual data like specific words. Alternatively, if distinction is only manageable for a complex model this would mean that the difference are more nuanced for example based on writing style.

Preprocessing of the textual data in the political articles is performed by removing punctuation, special characters and stop words. Furthermore, the text is converted to lowercase. Thereafter, for the non-transformer models, TF-IDF features are extracted using sklearn's TfidfVectorizer. A minimum document frequency of 30 is used to eliminate infrequent words to improve performance.

A baseline performance is set on the task by analysing the results of a dummy classifier, in the form of a model which randomly guesses the class independently of the input to the model, as well as a model which predicts the most frequent class.

Experimentation is performed using various modeling techniques: Decision Tree, RBF Support Vector, XGBoost, KNeighbors, Gaussian Naive Bayes, Multinomial Naïve Bayes, Linear Support Vector and RobBERT v2. The sklearn implementation is used except for XGBoost which has its own Python package and a Dutch BERT model [7] which is implemented through the HuggingFace Transformers package. From the political subset of articles 80% is used as training set and 20% as test set. The performance of the models is compared in terms of macro F1-score. For each of the non-transformer models, default parameters were used. For RobBERT v2 the parameters that were used are a learning rate of 1e-5, batch size 16, 3 training epochs and weight decay of 0.01.

## 3.4 Article Generation

Modern transformer models enable automatic natural language generation. It is possible to fine-tune these on specific source material to generate text in the style of an author or news medium. For sports articles automatic generation is already in use [17]. Could this be followed up by reporting objective news? Or drawn a step further even in analysis articles? If so, who is considered the author and who is responsible for what is written? Inversely, this technique could be used to analyse the general writing style or bias of an author or news medium. These models are a form of extrapolation of writing. Therefore, this analysis is limited by its assumption that this extrapolation is an accurate representation of how the author or news medium writes.

With respect to sub question 5, we fine-tune generative natural language models for each newspaper separately, and analyse and compare generated articles based on a common prompt. A pre-trained Dutch version of GPT-2 is utilised, GroNLP's small dutch model [29]. This model recycles the original English GPT-2 model [21]. The recycling means retraining the lexical embeddings of the originally English model for Dutch alternatives while fixing the transformer layers. This retraining of the lexical embeddings is performed with a dataset consisting of Wikipedia (2.8GB), newspaper articles(2.9GB) from 2007, books(6.5GB) and articles from various Dutch news websites(2.1GB). The model can be fine-tuned on a specific textual data set. In the experiments we use GroNLP's small dutch model zero-shot. Furthermore, the model is fine-tuned on the NRC, Volkskrant, Trouw and Het Parool political subset as well as these collectively, which is described as the general model.

The text generation model works on the basis of prompts. This is an initial piece of text from which the model can predict the next word to follow. Reiterating this process of predicting the next word results in the generation of a text.

The analysis of generated data offers advantages over analysing the source data. First, experiments can be performed very specifically due to the text being generated on the basis of a specific prompt. Second, as the textual data are generated on the basis of a language model, the samples can be considered a general collective style or writing angle of the complete source. For example, a single author from a newspaper may have a different style than all the writers in the newspaper combined. The disadvantage is that the analysis relies on the assumption that the generated textual data is a correct representation of the writing of the source.

First, the divergence in text generation of these models is compared through initiation of the different model versions with a neutral prompt. Second, samples from the different models on a specific subject are compared to analyse the extent to which the generated text diverge in terms of theme. Third, a larger sample is analysed in terms of sub-topics generated. Topic modeling is applied to find the extent to which the generated text contains distinguishable sub-topics. Finally, samples from the different models covering politicians and parties are evaluated in terms of sentiment to analyse the extent to which the overall sentiment diverges for each generative model.

# 4 Data

The data used to answer the research questions is collected by scraping articles from the Internet archives of various Dutch newspapers. As the second sub-question covers the shared political topics of newspapers, a broad scope of sources is required. Therefore, a balanced and representative collection is the key to establishing a suitable analogue of political discourse. However, some newspaper websites have restrictions on crawling and scraping activities. Therefore, the data are limited to articles from NRC (centrist, progressive liberal), Volkskrant (centre left, progressive), Het Parool (Amsterdam regional, centrist) and Trouw (centre, protestant origins).

## 4.1 Data Collection

The collection of articles for each newspaper is carried out following the same general sequence of steps. First, the website archive is crawled to index all the articles URLs over the course of the full year 2021. Second, all these links are scraped using Python's Requests library, resulting in a collection of HTML data for each web page. Third, the HTML data is parsed to produce clean text article data as well as publication timestamp, title and category information. This step is especially tricky, which resulted in some missing data points with respect to the publication timestamp and category. Capitalisation and punctuation are retained. Fourth, the data are saved as JSON dumps.

## 4.2 Results

In total 96840 articles have been collected, respectively for NRC (32043), Volkskrant (25702), Trouw (20944) and Parool (18151) as visualised in Figure 2. To illustrate the size of these

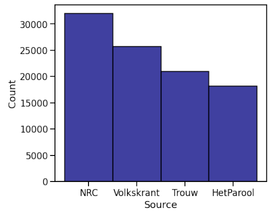


Figure 2: Total number of articles

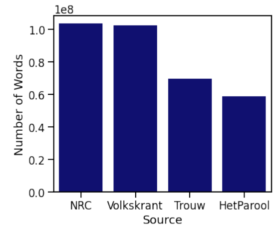


Figure 3: Total number of words (in 100Ms)

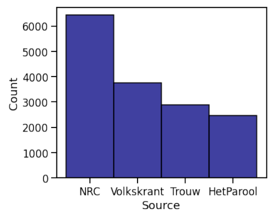


Figure 4: Number of articles in political subset

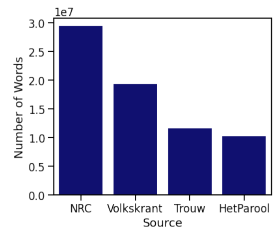


Figure 5: Number of words in political subset (in 10Ms)

collections the number of words in each set is illustrated in Figure 3. A subset of 15,508 articles is connected to politics through the mention of either a party leader or party name in the 2021 Tweede Kamer, the Dutch House of Representatives. This set consists of the articles from the complete set that contain either a party name or a party leader name. This political subset consists of 15.508 articles, respectively for NRC (6425), Volkskrant (3752), Trouw (2877) and Parool (2454) as visualised in 4. To illustrate the relative size, the number of words for each source is visualised in Figure 5.

Although the total number of articles in Volkskrant is significantly smaller than the number of articles in NRC the total number of words in these articles is comparable. Thus, Volkskrant writes less but longer articles. Furthermore, with regard to the political subset, NRC has a relatively larger number of articles and especially compared to the Volkskrant a larger number of words in articles concerning politics.

# 5 Experiments and Results

This section gives an overview of all the experiments and results carried out with the methods described in the Methods section 3.

## 5.1 Coverage Bias

In order to illustrate the political coverage bias present in newspapers the results of the Log Normalised Mention Frequency, as described in 3, of parties and politicians in 2021 are depicted in Figure 6 and 7. In both figures, the parties or politicians on the y-axis are ordered according to the party seats in the Tweede Kamer. Therefore, one would expect the coverage to gradually decrease from the biggest party at the top towards the smallest party at the bottom.

Two parties clearly break this idea: PVV and FVD both have contrasting low coverage compared to the other parties. Both are considered (far) right wing populist parties, which

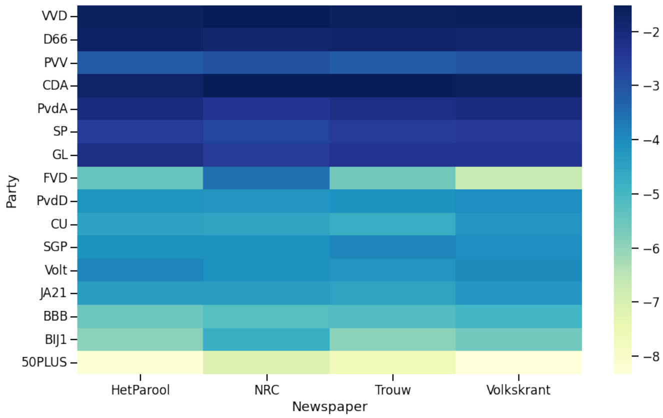


Figure 6: Party-newspaper coverage

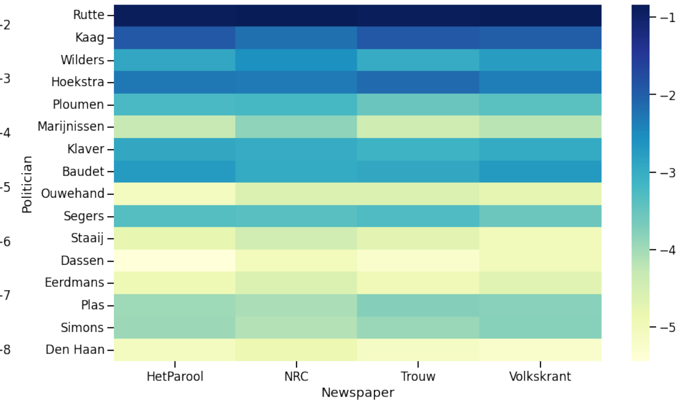


Figure 7: Politician-newspaper coverage

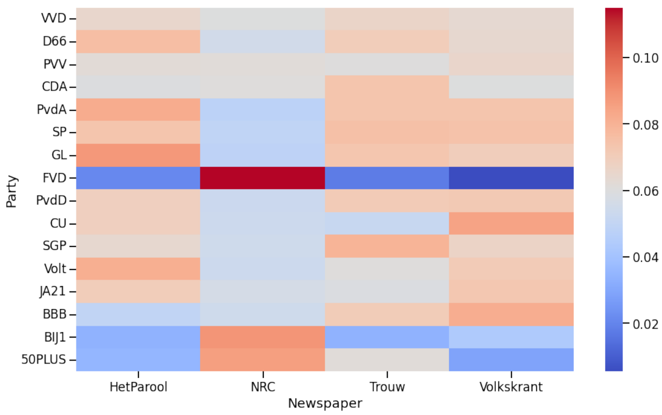


Figure 8: Relative party-newspaper coverage

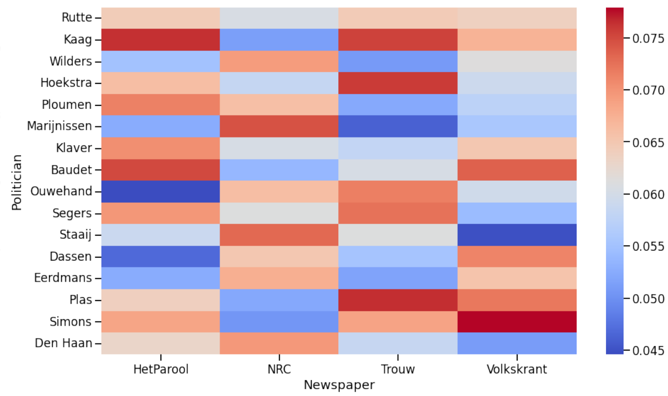


Figure 9: Relative politician-newspaper coverage

may explain this discrepancy. Interestingly, the low coverage of PVV and FVD contrasts with relatively regular coverage of the party leaders Wilders and Baudet, with Baudet scoring better in relative terms. This could alternatively explain the lower mention frequency of the parties as the party leader is mentioned instead. With respect to figure 7, the odd one out is Marijnissen, although her party (SP) has the same amount of seats as Ploumen's PvdA, she is mentioned less overall. Furthermore, the relatively high coverage of Segers could be attributed to the fact that his party was part of the government.

With an increased contrast, the Relative Normalised Mention Frequency, as described in 3 is depicted in Figure 8 and 9. With respect to parties, no large differences in coverage are seen, except for the FVD which is mentioned significantly more in NRC and less in Het Parool, Trouw and Volkskrant. Apart from Mark Rutte who is covered fairly consistent over all sources, the contrasts in politician coverage are more prevalent. Kaag for example, does relatively well in Het Parool and Trouw. Hoekstra and van der Plas are relatively prominent in Trouw and Simons in Volkskrant.

In conclusion, the results for the Log Normalised Frequency and Relative Normalised Frequency show disproportionate coverage of certain politicians and parties. Which indicates a certain bias in news coverage.

## 5.2 General and Source Specific Political Discourse

In order to analyse the shared political topics and source specific topics in newspapers topic modelling and word embeddings are used as described in 3.

### 5.2.1 Topic Modelling

The results of the LDA topic modelling in the political subset are described in Table 2. The assigned topics are ordered in marginal topic distribution. This can be interpreted as the importance of a topic with respect to the corpus.

The most prominent topic is national politics and corona policy. Furthermore, topics consisting of far right/left, EU, and international politics are distinguished. Finally, thematic topics on family life, law and order, economy, elections and personal assets/debt are present. One of the topics has not generalised to an interpretable topic or theme and thus is left blank. The first four clearly political topics are the most prominent; it is interesting to see which additional politically related topics arise. These themes can give insight into the topics discussed in a political context. Thus, illustrating the agenda setting in general political discourse.

The source specific topic modeling is analysed with regards to the topic modeling results on the general political discourse based on the complete collection of articles. The topics or themes that arise in the modelling of articles from a particular source are considered a subset of the general political discourse. Following is an analysis of the topics for each source respectively.

With regard to the NRC, topics follow the general political discourse with the two most prominent topics being: national politics/corona policy and far left/right politics. However, the third and an additional eighth topic in the newspaper cover national politics with regard to distrust of government. Several affairs which depicted governmental shortcoming arise. Furthermore, themes of family life, antivax movement and corona/asylum seekers are present.

Regarding Trouw, the most prominent topic consists of national politics, followed by government formation. Corona policy is less prominent compared to the NRC and all other topics consist of national politics with one of them related to distrust of government.

With regard to Het Parool, this Amsterdam-focused newspaper follows the trend of having national politics as the most prominent topic. Furthermore, a far-right/left political topic arises, and two topics on Amsterdam specific politics.

With regard to the Volkskrant, the results consist of very clearly distinguishable topics. Again, national politics are most prominent. Furthermore, the political topics corona policy, far right politics, distrust of government and far left arise. In addition to themes such as life, culture and living/work.

In conclusion, national politics is the only topic consistently found for all the sources. Thus, a limited overlap of political topics is present based on this analysis.

### 5.2.2 Word embeddings

Representing the textual data in vector space and visualising the respective embeddings of parties and politicians in a lower-dimensional space gives an intuition to how is written about parties or politicians in general or source-specific.

The text data is preprocessed by removing punctuation and special characters as well as lowercasing the text.

| | General | NRC | Trouw | Het Parool | Volkskrant |
|---|---|---|---|---|---|
| 1 | Domestic Policy, Corona Policy | Domestic Policy, Corona Policy | Domestic Policy | Domestic Policy | Domestic Policy |
| 2 | Far Right/ Left | Far Right/ Left, Purchasing Power | - | - | Corona Policy |
| 3 | EU Politics | (Distrust) Domestic Politics | Coalition Formation | Domestic Policy | Life |
| 4 | Foreign Politics | Family Matters/Housing/Living | - | Far Right/ Left | Far Right |
| 5 | Family Matters | 'Wappies' (Corona conspiracy) | - | - | (Distrust) Domestic Politics |
| 6 | Safety and Law Enforcement | - | Domestic Policy/ Corona Policy | - | (Far) Left |
| 7 | - | Corona/ AZC (refugee centers) | Domestic Policy | Amsterdam Politics | Culture |
| 8 | Economy | (Distrust) Domestic Politics | Domestic Policy | - | Housing and Work |
| 9 | Elections | - | (Distrust)Domestic Politics | - | Safety and Law Enforcement |
| 10 | Wealth/Debt Management | - | Domestic Policy | Amsterdam Politics | - |

Table 2: LDA Topic Modelling (translated into English)

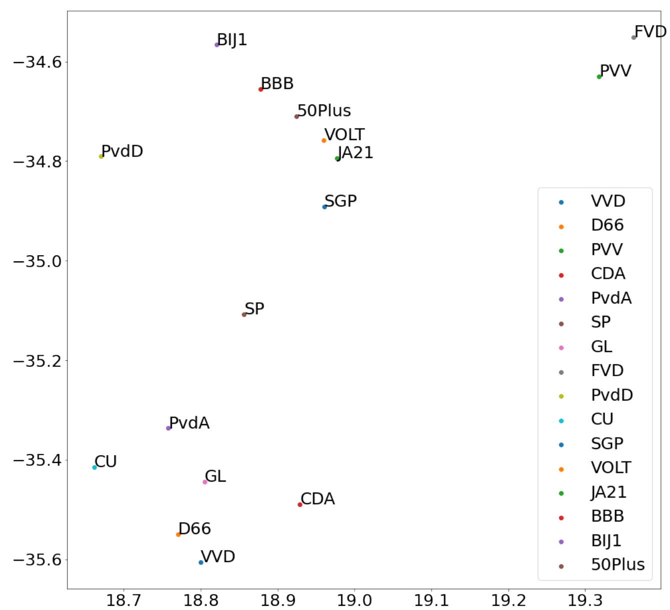


Figure 10: Word2Vec + tSNE: Parties in general

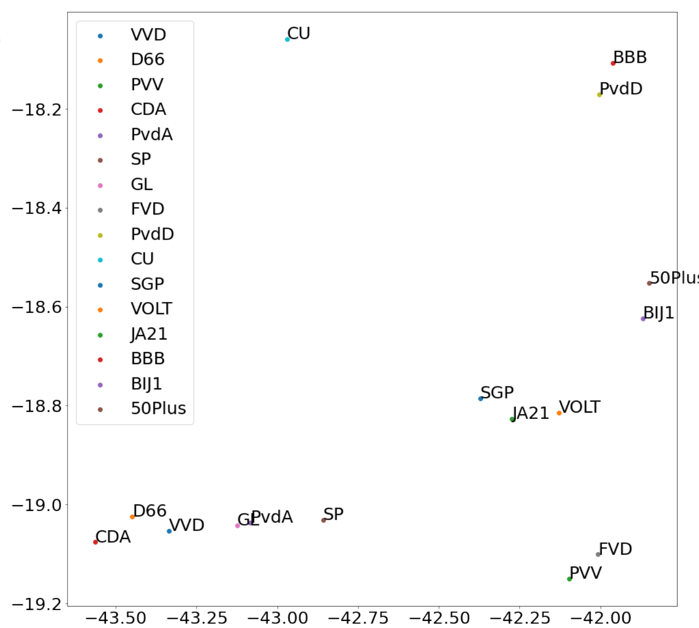


Figure 11: Word2Vec + tSNE: Parties in NRC

The general political discourse is visualised in Figure 10. The grouping of the parties that end up in government in 2022 is distinguished at the bottom left. VVD, CDA, D66 and CU. GL, PvdA and SP are also in the vicinity which may be explained by their efforts to be part of the formation. The farthest away from this governing party group we find the FVD and PVV in the top right. These parties are both considered far right and therefore may profile themselves opposing the established parties. The remaining parties can be described as the moderate opposition.

The source-specific political discourse is visualised in Figure 11 up to 14. Concerning the NRC figure, a similar grouping of governing parties is present in the bottom left of the figure along with GL, PvdA and SP in the vicinity. The CU is located far away at the top of the figure. The NRC mentions the CU in a relative distant context from the governing parties. Moderate opposition parties are found on the middle right. Compared to the general political discourse visualised in Figure 10, the most distant parties from the governing parties are now BBB and PvdD in the upper right. The far-right parties FVD and PVV, located on the bottom right, are relatively close to the governing and moderate opposition parties.

| Model | Accuracy | F1-score | | Model | Accuracy | F1-score |
|---|---|---|---|---|---|---|
| Majority Class | 0.41 | 0.14 | | KNeighbors | 0.42 | 0.35 |
| Random Guess | 0.24 | 0.23 | | Gaussian Naive Bayes | 0.40 | 0.38 |
| Decision Tree | 0.43 | 0.27 | | Multinomial Naïve Bayes | 0.47 | 0.33 |
| RBF Support Vector | 0.55 | 0.46 | | Linear Support Vector | 0.53 | 0.48 |
| XGBoost | 0.51 | 0.41 | | RobBERT | **0.87** | **0.86** |

Table 3: Performance of source classification models

Regarding the Volkskrant, there is a grouping of the governing parties together with GL and CU. Yet, CU again finds itself singled out away from this group. The opposition is located in the vicinity and among them PVV and FVD are present. Which is interesting as in NRC and the general embedding space took their own space. Furthermore, the smallest parties are in the distant topright: BBB, 50Plus and Bij1. They are interestingly joined by PvdD which has many more seats.

With regard to Trouw, most parties are clustered on top of each other in the bottom left. CU and BBB are located in the middle between this large group and 50Plus, Bij1 and PvdD are located on the far right in the embedding space. Interestingly here again FVD and PVV are part or in the vicinity of the governing parties.

With regard to Het Parool, the governing parties cluster in the top left with CU in the vicinity. PvdA, GL, SP and PVV are in this cluster as well. A second group closeby is formed by CU, SGP, FVD and JA21. 50Plus, BBB, PvdD and Bij1 are distant in the right of the embedding space and Volt is distant in the bottom left. Unique for Het Parool, as usually it finds a place amongst the moderate opposition parties.

In conclusion, these results do not show a consistent shape of the parties in the embedding space. This is an indication against the presence generally shared political discourse. Further visualisations of the word embeddings of politicians in general and source specific political discourse is given in section 9.

## 5.3 Discriminating Newspapers by Article Texts

For the non-transformer models TF-IDF features are extracted using sklearn's TfidfVectorizer. A minimum document frequency of 30 is used to eliminate infrequent words to improve performance. This results in a vocabulary of 11804 words. This minimum document frequency is used to prevent the vocabulary from having a unmanageable size.

The performance of each of the models applied to the source classification task is given in table 3. When comparing the simpler models (Decision Tree, KNeighbors, Gaussian Naive Bayes, Multinomial Naive Bayes) with the Majority Class dummy classifier only a small improvement in accuracy is seen, though the F1-score does get improved significantly. Runner up are the RBF Support Vector, Linear Support Vector and XGBoost models. They show a significant improvement in F1-score and an accuracy of >50%. The best performance is found with the most advanced model, RobBERT.

With respect to the linear SVC model the importance of features can be interpreted by analysing the size of the coefficients of the one-vs-one classifiers. For each class combination, the top ten positive and negative predictors are visualised. With respect to Parool-NRC, Figure 15, it is logical to see 'amsterdam' as a strong positive feature and 'nrc' as a strong negative. With respect to Parool-Trouw in Figure 16, it is interesting to see 'mark' a strong predictor

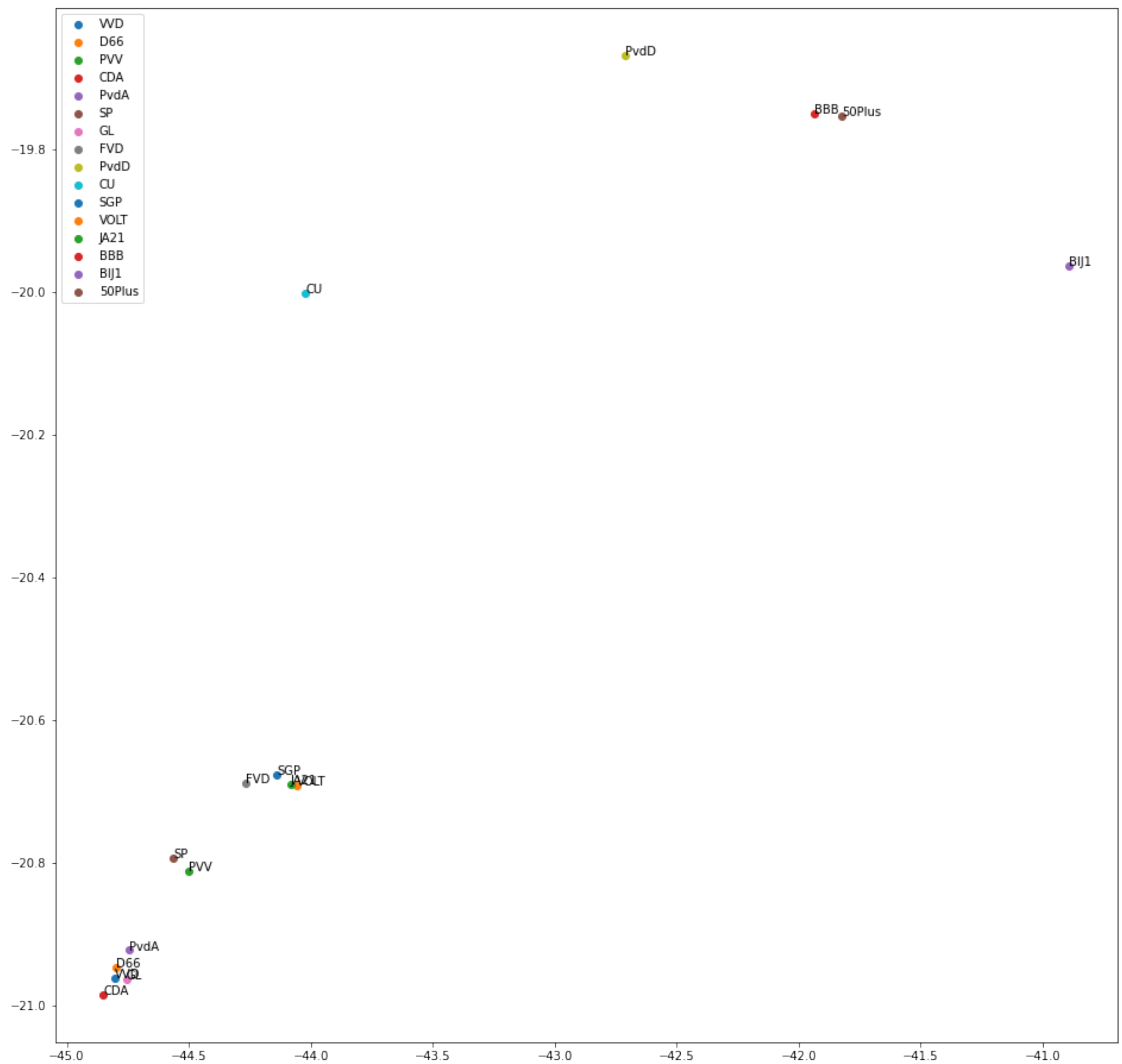


Figure 12: Word2Vec + tSNE: Parties in Volkskrant

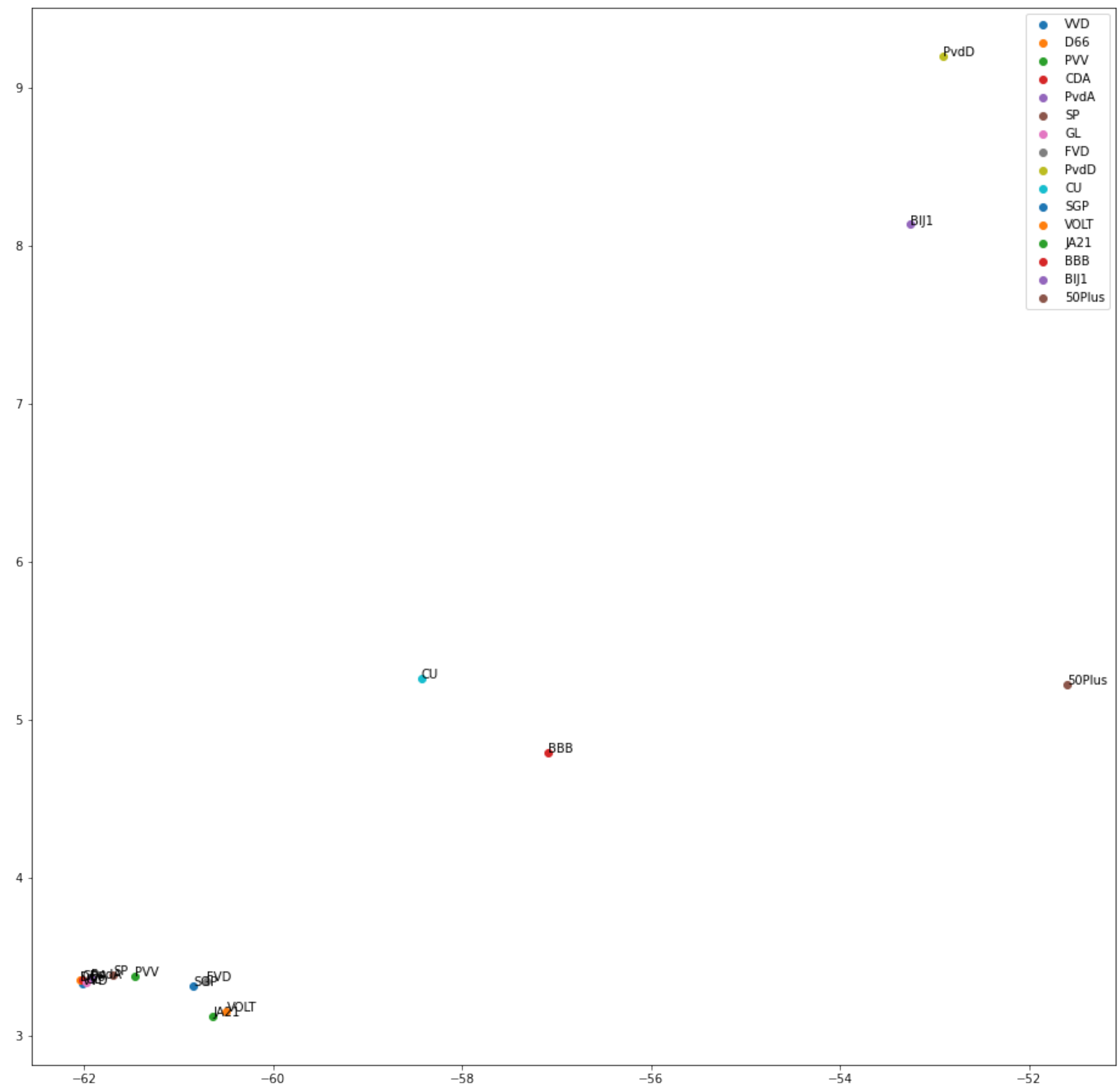


Figure 13: Word2Vec + tSNE: Parties in Trouw

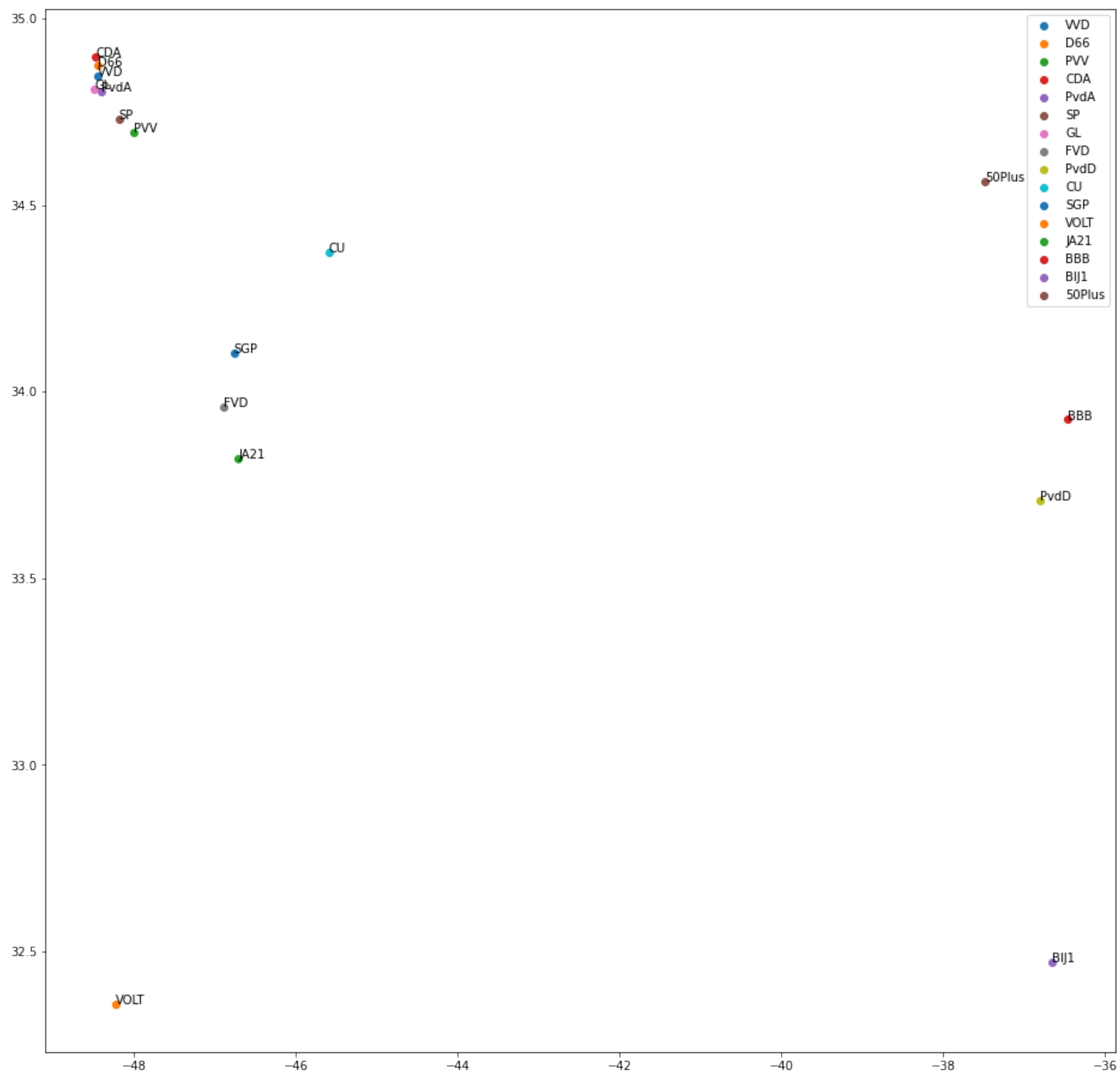


Figure 14: Word2Vec + tSNE: Parties in Het Parool

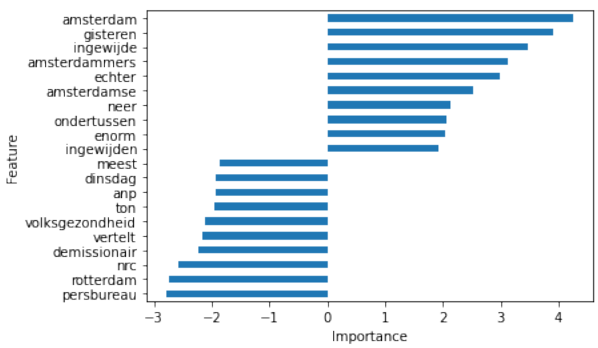


Figure 15: Parool vs NRC

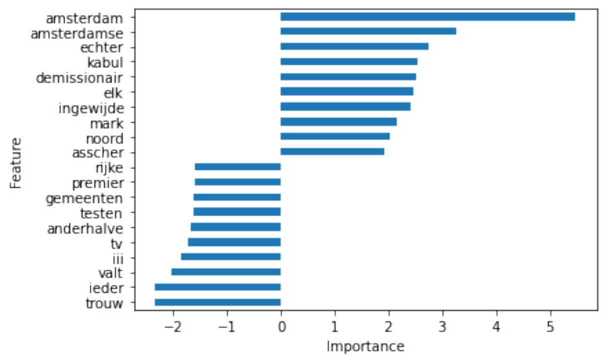


Figure 16: Parool vs Trouw

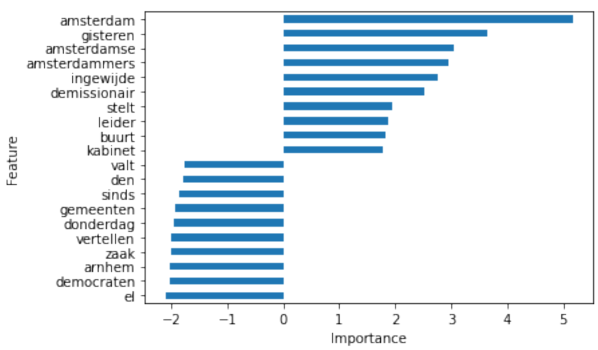


Figure 17: Parool vs Volkskrant

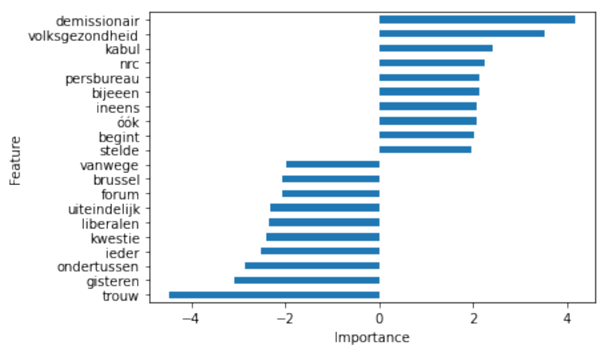


Figure 18: NRC vs Trouw

for Parool in contrast with 'premier' for Trouw. With respect to Parool-Volkskrant, Figure 17, Parool shows logical predictors like 'amsterdam'. Yet, the Volkskrant predictors are not very explicitly typical for the newspaper from a human perspective. With respect to NRC-Trouw, Figure 17, 'trouw' and 'nrc' are logical predictors yet, 'BIJ1' for NRC and 'forum', referring to FVD, are striking features. With respect to NRC-Volkskrant, Figure 19, 'nrc' and 'volkskrant' are logical. Yet, 'BIJ1' and 'Sywert' for NRC are interesting as they are respectively a political party and a party ideologist for CDA. Finally, with respect to Trouw-Volkskrant, Figure 20, 'trouw' is a very logical predictor for Trouw. In contrast, the predictors for Volkskrant are again not explicitly linked to the newspaper from a human perspective.

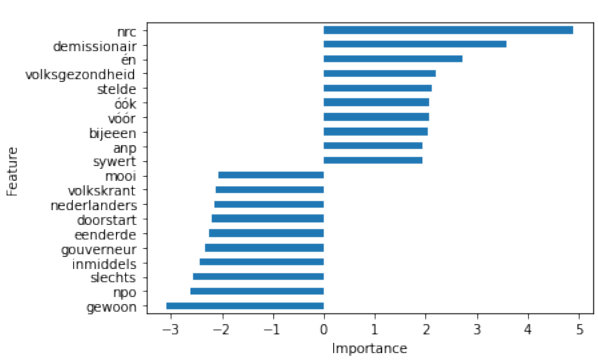


Figure 19: NRC vs Volkskrant

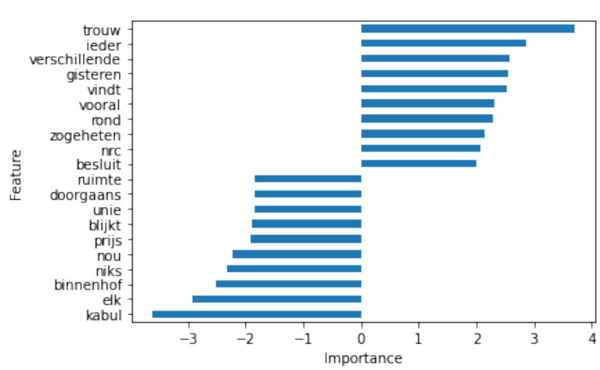


Figure 20: Trouw vs Volkskrant


Figure 21: General Wordcloud


Figure 22: NRC Wordcloud

## 5.4 Article Generation

The model is implemented using the Huggingface's transformer package. Specifically, the following subpackages are used AutoTokenizer, TextDataset, DataCollatorForLanguageModeling, Trainer, TrainingArguments and AutoModelWithLMHead are used. The model is fine-tuned using the full collection of articles as described in Section 4. A maximal sequence length of 128 tokens was used with truncation, a batch size of 32, prediction loss only and a warm up of 500 steps for the learning rate scheduler.

This section is structured in the following analyses: Neutral Prompt-, Thematic-, Subtopic- and Sentiment Divergence.

### 5.4.1 Neutral Prompt Divergence

In order to analyse the generative models with a neutral prompt, sampling is performed based on: "X houdt een toespraak". This translates to: "X gives a speech". A 1000 samples are produced with the maximal sequence length set to 30. From the collective produced text, word clouds are produced where the most prominent terms are displayed scaled to there occurrence using the Wordcloud python package. For this visualisation the words "X", "gives", "a" and "speech" are removed from the samples.

Two of the generated samples are:

- "X houdt een toespraak op vrijdag 17 maart over het coronabeleid. Op dat moment zijn in totaal 2.870 mensen ingeënt, blijkt uit de voorlopige cijfers van het Centraal Bureau voor de Statistiek (CBS). Van hen liggen er 1."
- "X houdt een toespraak van premier Mark Rutte en Tweede Kamerlid Pieter Omtzigt over het klimaatbeleid. We moeten ons afvragen wat we doen als er niet meer wordt geïnvesteerd in hernieuwbare energie, zoals fossiele brandstoffen, aldus Omtzigt"

The word cloud based on the general text generator model, Figure 21, prominently contains two names of politicians, Rutte and de Jonge. During the corona crisis they gave speeches together informing the public of corona measures. Considering the NRC-word cloud, Rutte and D66 are very prominent and the other words in the cloud cover corona measures and infections. The Volkskrant-word cloud interestingly does not show Rutte yet it does contain "president", which could refer to him. D66 does appear yet, it is quite small. Furthermore, not much reference to the corona crisis is present. The Trouw word cloud interestingly diverges towards the 2021 inaugurated president Joe Biden of the United States. These results show a significant divergence in topics resulting from a neutral prompt.


Figure 23: Volkskrant Wordcloud


Figure 24: Trouw Wordcloud

| *Source* | *Sample* |
|---|---|
| *NRC* | *'Abortus is in Nederland een strafbaar feit. Dat heeft de Tweede Kamer deze vrijdag besloten, meldt het demissionair kabinet. Het besluit komt neer op'* |
| *Volkskrant* | *'Abortus is in Nederland verboden, omdat de foetus niet meer zwanger wordt van zijn of haar biologische moeder. Dat heeft het ministerie van Volksgezondheid dinsdag bekendgemaakt'* |
| *Trouw* | *'Abortus is in Nederland een onderwerp van veel kritiek. De afgelopen jaren zijn er talloze onderzoeken gedaan naar het gebruik van abortus, onder meer bij vrouwen die'* |
| *Het Parool* | *'Abortus is in Nederland de enige vorm van staatsrecht. Het recht op abortus heeft betrekking op het zelfbeschadiging van kinderen, en niet op'* |

Table 4: Samples based on thematic prompt

### 5.4.2 Thematic Divergence

In order to analyse the thematic divergence ten samples of length 30 are produced on a controversial topic and analysed manually. The prompt "Abortus in Nederland" is used which translates to "Abortion in the Netherlands". The NRC generative model produces text on the themes of forbidden, (il)legal and punishable. The Trouw generative model produces text on the themes of discussion, women and pregnant. The Volkskrant generative model produces text on the themes of forbidden, important topic and women. The Parool generative model produces text on the themes of (il)legal, rights and Nederland. This shows a thematic divergence as all the generative models approach the topic from a different angle. Some samples are given in Table 5.4.2.

### 5.4.3 Subtopic Divergence

In order to analyse the subtopic divergence of generated text a sample is taken from the generative model fine-tuned on all the newspapers. The prompt which is used is "Het corona beleid" which translates to "The corona policy". The generated length of the samples is 100 tokens and the number of samples is 500. This results in a total collection of 50k words. This sample is analysed using LDA topic modelling to 5 topics which is visualised in Figure 25. The results show the model has produced a wide range of terms and names involved with corona policy, yet the resulting LDA topics are not distinguishable to a specific theme. So the

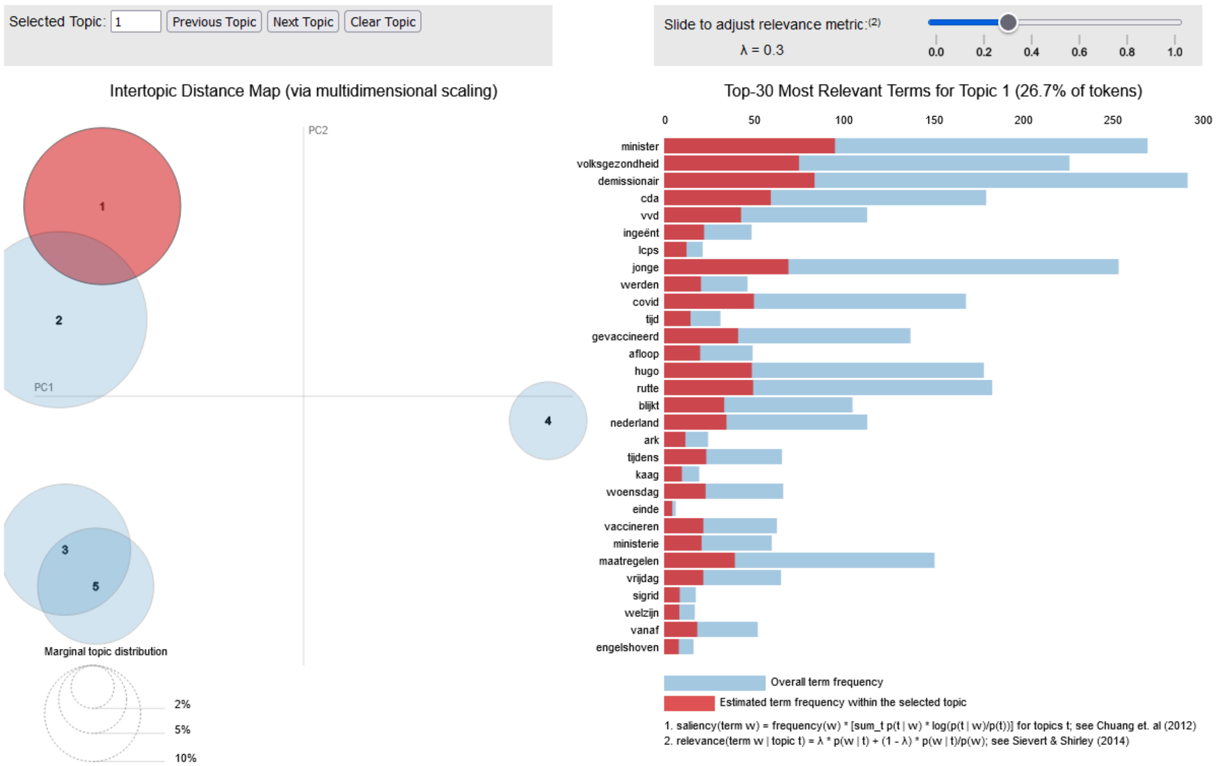


Figure 25: LDA "Het corona beleid"

divergence in terms of sub-topics of "The corona policy" is not consistent.

### 5.4.4 Sentiment Divergence

In order to analyse the divergence in sentiment toward politicians or parties for each of the generative models samples are taken and evaluated on their sentiment. For this sentiment evaluation NLP Town's bert-base-multilingual-uncased model is used which is fine-tuned for sentiment analysis in six languages with a Dutch training subset of 80k reviews. This model predicts the sentiment on a 1-5 rating. For each specific prompt, 50 samples are taken from each of the generative models with a sequence length of 50 tokens. The prompt which is used is "X staat bekend om" which translates to "X is known for". Here X is replaced by a party or politician.

The results are visualised in histograms. Figure 26 shows divergence in the sense of relative large amount of positive sentiment scoring for Mark Rutte in Het Parool and a relative little amount of negative sentiment scoring in NRC. Figure 27 shows an overall more positive sentiment. For Het Parool and Trouw the majority of text is evaluated with a sentiment score of 4 and Volkskrant a large amount of positive sentiment score of 4 as well. Figure 28 shows a similar positive shape for Trouw and Volkskrant. However, for NRC and Het Parool the majority of text is evaluated with a neutral score of 3. Furthermore, in Figure 29, an overall negative sentiment is seen. A significant amount of text for all generative models is evaluated with a sentiment score of 1. From all models, Trouw is relatively the most neutral towards Thierry Baudet.

The results show significant differences in the sentiment toward specific politicians for each of the source specific trained models. Thus, a divergence of sentiment is present to a certain

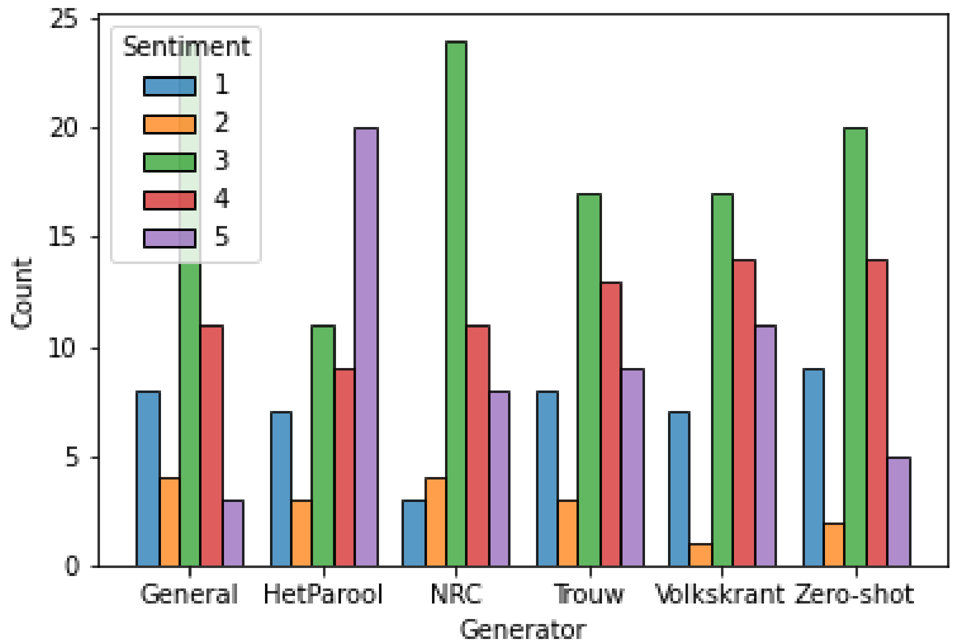


Figure 26: Sentiment analysis generative models Mark Rutte

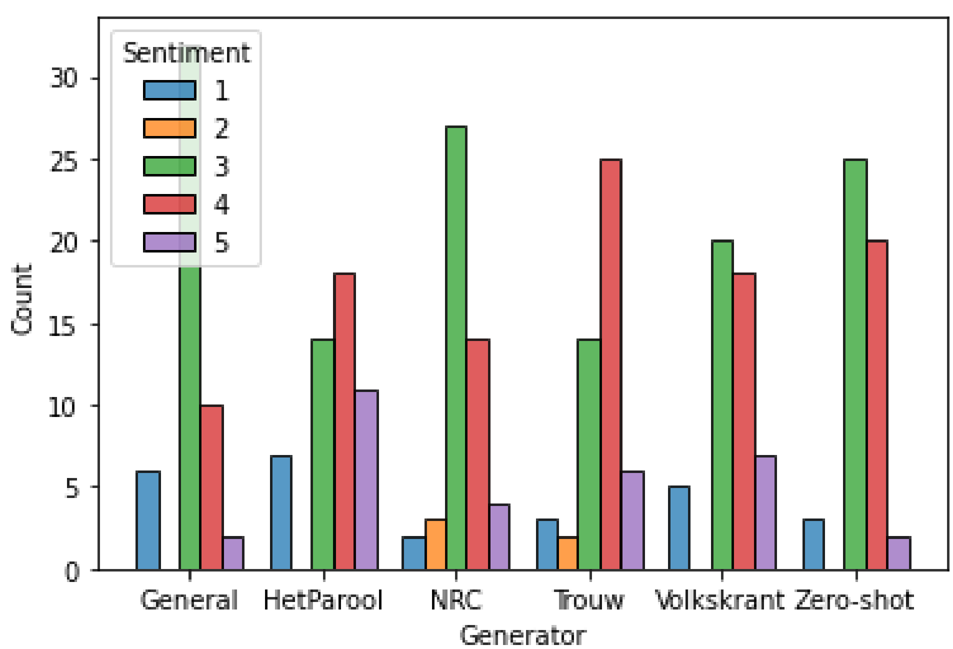


Figure 27: Sentiment analysis generative models Sigrid Kaag

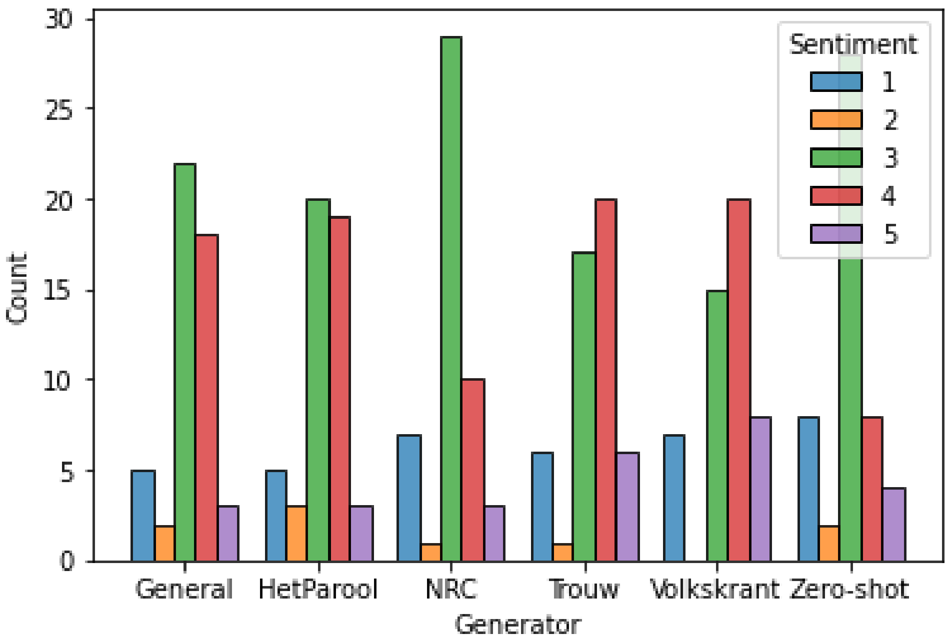


Figure 28: Sentiment analysis generative models Wopke Hoekstra

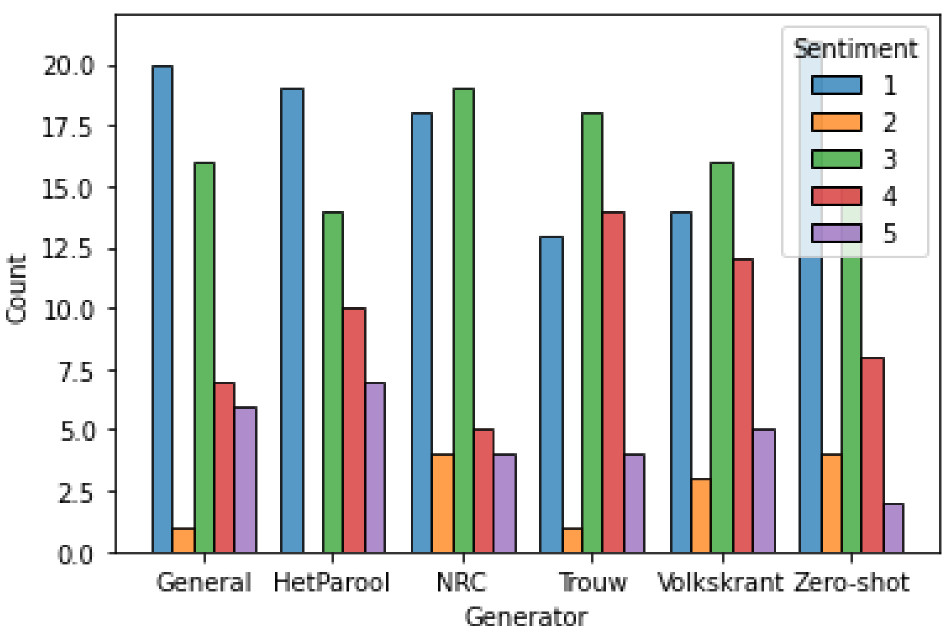


Figure 29: Sentiment analysis generative models Thierry Baudet

extent. Further analysis of the party sentiment divergence is available in section 9.

# 6 Discussion

Concerning SQ1, on political bias, some results stand out. A low coverage in relation to the respective number of seats of far right parties is present in three newspapers. With respect to the politicians there are four that receive an unexpected amount of coverage either too high or low, including the far right. This indicates that there is a bias present in Dutch newspapers in terms of coverage.

Concerning SQ2, on general political discourse, a set of clear topics is distinguished in the collective of newspapers through LDA visualisation analysis. Furthermore, representation of the political articles in vector space results in a structured clustering of political parties in government, opposition and far right.

Concerning SQ3, on newspaper-specific political discourse, the most prominent topic of national politics is shared consistently by all newspapers. However, apart from this topic, the newspapers differ significantly. It should be taken into account that the topics for the newspaper specific LDA analysis were much less distinguishable. This is probably due to the smaller size of the subsets. With regards to the vector space representation of the newspaper-specific political discourse, the structure of the parties is comparable to the general political discourse cluster-wise: governing, far right and moderate opposition. However, how these clusters are located in the space is considerably different. Thus, taking into account the topic modelling and vector representation analysis, the newspaper-specific political discourse differs considerably from the general political discourse.

For a subset of topics in the LDA analysis it is difficult to distinguish a clear theme or subject. These are left as – in table 2. This is primarily the case for Trouw and Het Parool and probably so as these collections of articles are significantly smaller, see Section 4, than NRC and Volkskrant where topics do generalise to well distinguishable subjects or themes.

Concerning SQ4, on the identification of a newspaper given an article, this task was very manageable for the advanced RobBERT model reaching high accuracy and F1-score. For simpler methods some learning or performance was reached yet no accuracy higher than 0.55 was reached. Thus, identifying the newspaper for which an article was written is a complex but achievable task.

Concerning SQ5, on the divergence of text generation models trained on specific newspapers, each of the fine-tuned models takes its own direction when prompted neutrally, on a controversial topic and when evaluated for sentiment.

The key advantages off applying generative AI in this context is that experiments can be performed very specifically due to the text being generated on the basis of a specific prompt. Furthermore, as the generated text is based on a language model it can bee seen as a more general representation of the writing style of an author or source. These are just very simple initial experiments for illustrative purposes, but in our view already demonstrates that generative models can be interesting tools in this context, though one needs to take into account that this type of research is more speculative as it based on generated, synthetic data.

# 7 Research Limitations and Future Work

The analysis in this thesis rely on the dataset of articles that have been collected. Due to some newspaper websites disallowing crawling or scraping activities they could not be added to the research data set. It would have been interesting to incorporate a tabloid newspaper like the Telegraaf, a financial oriented newspaper like Financieel Dagblad and AD which characterises itself as politically and religiously neutral. Furthermore, the data used for this thesis is limited to the year 2021 and temporal effects are not analysed. For example, it would be interesting to train generative AI on data for each year from 2012 up to 2022 and analyse the sentiment divergence towards a politician or party.

# 8 Conclusion

In this thesis, research was carried out to analyse the extent to which the ideological position of newspapers is reflected in their writing. This subject was approached from several angles: measuring coverage bias, comparing general- and source specific discourse, performing classification of articles and analysing newspaper trained text generative models. The results showed several indications of political tendencies: disproportionate coverage of certain politicians and parties, limited overlap of political discourse, classifiable article source and divergence of generated text thematically and in terms of sentiment. Even though it is generally known that newspapers have their respective political ideology and are deemed to write accordingly, solely perceiving their writing on a left-to-right scale is inadequate as the political tendencies of newspaper are intricate. One should consider this when consuming from these media sources, and as in our new analysis, use a multitude of tools to analyze the data from multiple perspectives.

# References


[1] Abduljaber, M.: The dimensionality, type, and structure of political ideology on the political party level in the $arab$ world. Chinese Political Science Review **3**(4), 464–494 (2018)

[2] Abuhamad, M., Rhim, J.s., AbuHmed, T., Ullah, S., Kang, S., Nyang, D.: Code authorship identification using convolutional neural networks. Future Generation Computer Systems **95**, 104–115 (2019)

[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. the Journal of Machine Learning research **3**, 993–1022 (2003)

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[5] Coppock, A., Ekins, E., Kirby, D., et al.: The long-lasting effects of newspaper op-eds on public opinion. Quarterly Journal of Political Science **13**(1), 59–87 (2018)

[6] Dallmann, A., Lemmerich, F., Zoller, D., Hotho, A.: Media bias in German online newspapers. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media. pp. 133–137 (2015)

[7] Delobelle, P., Winters, T., Berendt, B.: RobBERT: a Dutch RoBERTa-based Language Model. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3255–3265. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.292

[8] Eberl, J.M., Wagner, M., Boomgaarden, H.G.: Are perceptions of candidate traits shaped by the media? The effects of three types of media bias. The International Journal of Press/Politics **22**(1), 111–132 (2017)

[9] Enevoldsen, K.C., Hansen, L.: Analysing political biases in Danish newspapers using sentiment analysis. Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift **2**(2), 87–98 (2017)

[10] Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research **61**, 65–170 (2018)

[11] Hassell, H.J., Miles, M.R., Reuning, K.: Does the ideology of the newsroom affect the provision of media slant? Political Communication **39**(2), 184–201 (2022)

[12] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (2019)

[13] Kang, H., Yang, J.: Quantifying perceived political bias of newspapers through a document classification technique. Journal of Quantitative Linguistics pp. 1–24 (2020)

[14] Kats, T., van der Putten, P., Schelling, J.: Distinguishing commercial from editorial content in news. Preproceedings 33rd Benelux Conference on Artificial Intelligence and the 30th Belgian Dutch Conference on Machine Learning (BNAIC/BENELEARN 2021), Luxembourg, November 10-12, 2021 (2021)

[15] Kawintiranon, K., Singh, L.: Polibertweet: A pre-trained language model for analyzing political content on twitter

[16] Kirk, H.R., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., Asano, Y., et al.: Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Advances in Neural Information Processing Systems **34**, 2611–2624 (2021)

[17] Kunert, J.: Automation in sports reporting: Strategies of data providers, software providers, and media outlets. Media and Communication **8**(3), 5–15 (2020)

[18] Lewenberg, Y., Bachrach, Y., Bordeaux, L., Kohli, P.: Political dimensionality estimation using a probabilistic graphical model. In: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. p. 447–456. UAI'16, AUAI Press, Arlington, Virginia, USA (2016)

[19] Liu, R., Jia, C., Wei, J., Xu, G., Vosoughi, S.: Quantifying and alleviating political bias in language models. Artificial Intelligence **304**, 103654 (2022)

[20] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[21] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[22] Ressmeyer, R., Masling, S., Liao, M.: "deep faking" political twitter using transfe r learning and gpt-2 (2019)

[23] Romanov, A., Kurtukova, A., Shelupanov, A., Fedotova, A., Goncharov, V.: Authorship identification of a Russian-language text using support vector machine and deep neural networks. Future Internet Journal **13**(1), 3 (2020)

[24] Schelling, J., van Eekelen, N., van Veelen, I., van Hees, M., van der Putten, P.: Bursting the bubble (extended abstract). In: MISDOOM 2020. p. 72 (10 2020)

[25] Sievert, C., Shirley, K.: LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. pp. 63–70 (2014)

[26] Susánszky, P., Kopper, Á., Zsigó, F.T.: Media framing of political protests–reporting bias and the discrediting of political activism. Post-Soviet Affairs pp. 1–17 (2022)

[27] Tangian, A.: Visualizing the political spectrum of Germany by contiguously ordering the party policy profiles. Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics **3**, 193–208 (2019)

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[29] de Vries, W., Nissim, M.: As good as new. how to successfully recycle English GPT-2 to make models for other languages. CoRR **abs/2012.05628**, 836–864 (2020), https://arxiv.org/abs/2012.05628

# 9 Appendix

## 9.1 Word Embeddings

Further visualisations as mentioned in 5. With regard to the visualisation of the vectors of politicians, in Figure 30 up to 34, the location of Kaag closeby Rutte in the general discourse stands out. For each of the source-specific discourses the distance between Rutte-Kaag and Rutte-Hoekstra is the same, yet collectively a closer relation appears. Furthermore with respect to the far right party leaders: Wilders and Baudet. Their location does not stand out in general discourse, Volkskrant, Trouw and Het Parool yet for NRC they do appear grouped away from other party leaders.

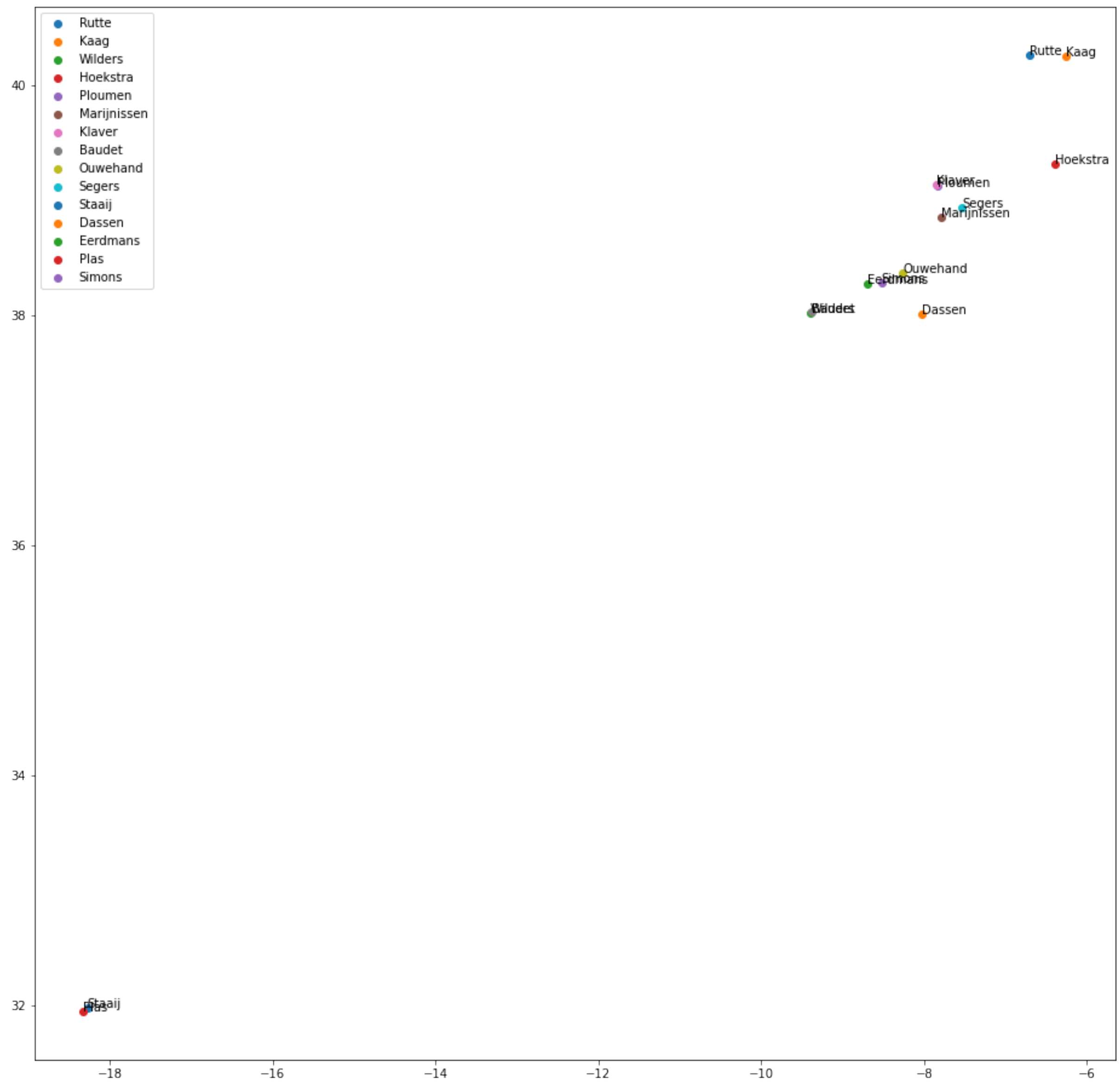


Figure 30: Word2Vec + tSNE: Politicians in General

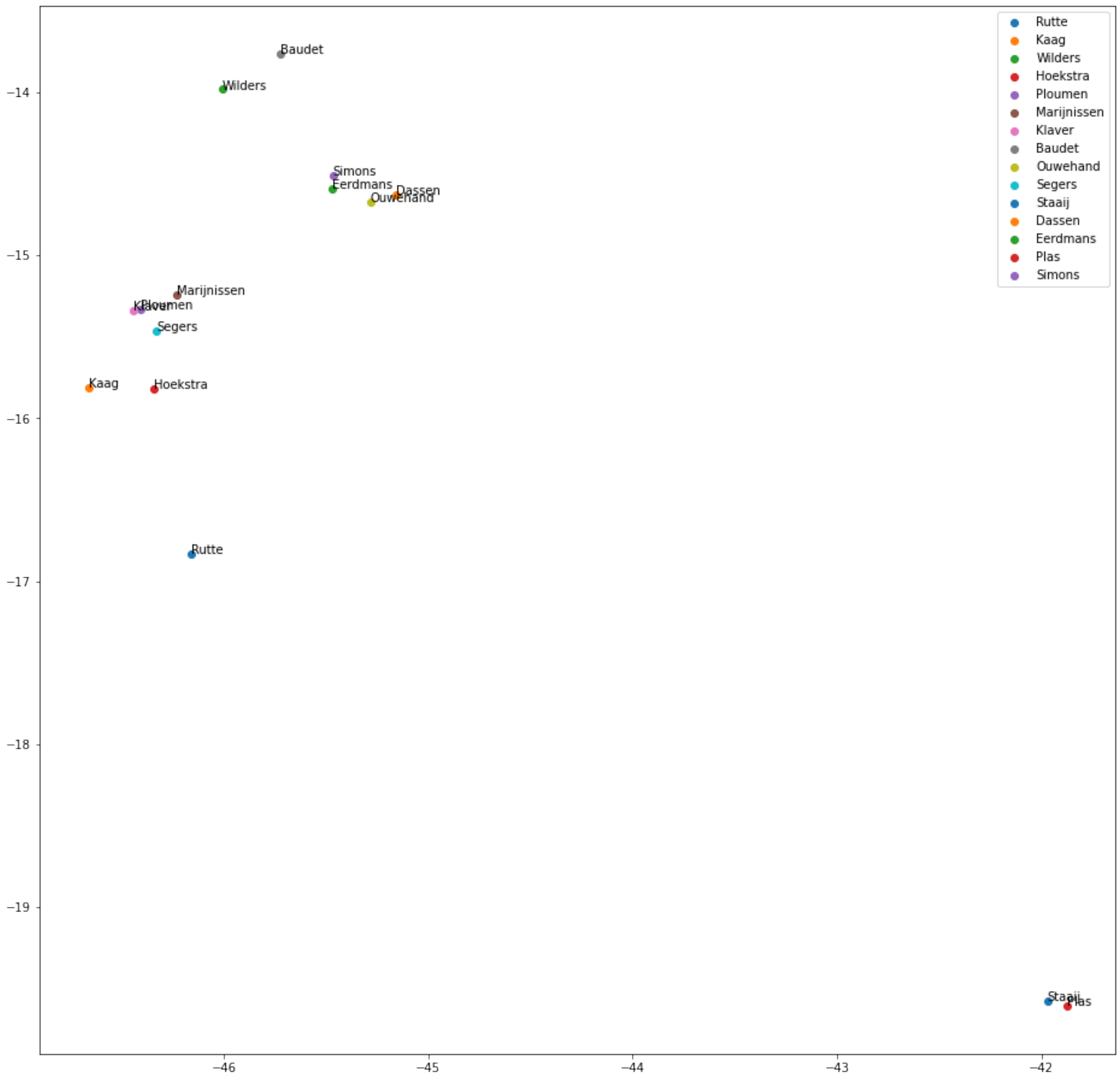


Figure 31: Word2Vec + tSNE: Politicians in NRC

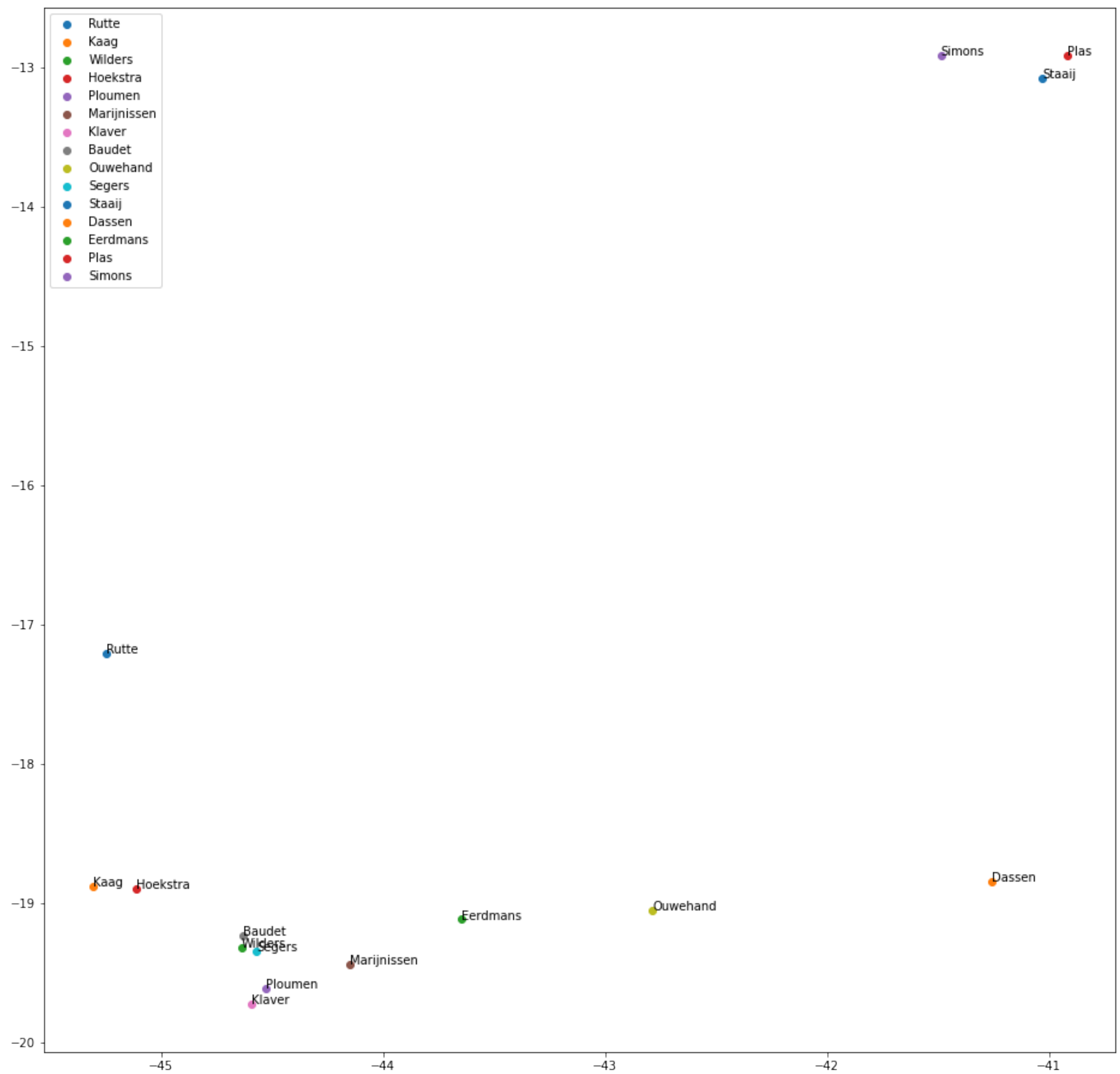


Figure 32: Word2Vec + tSNE: Politicians in Volkskrant

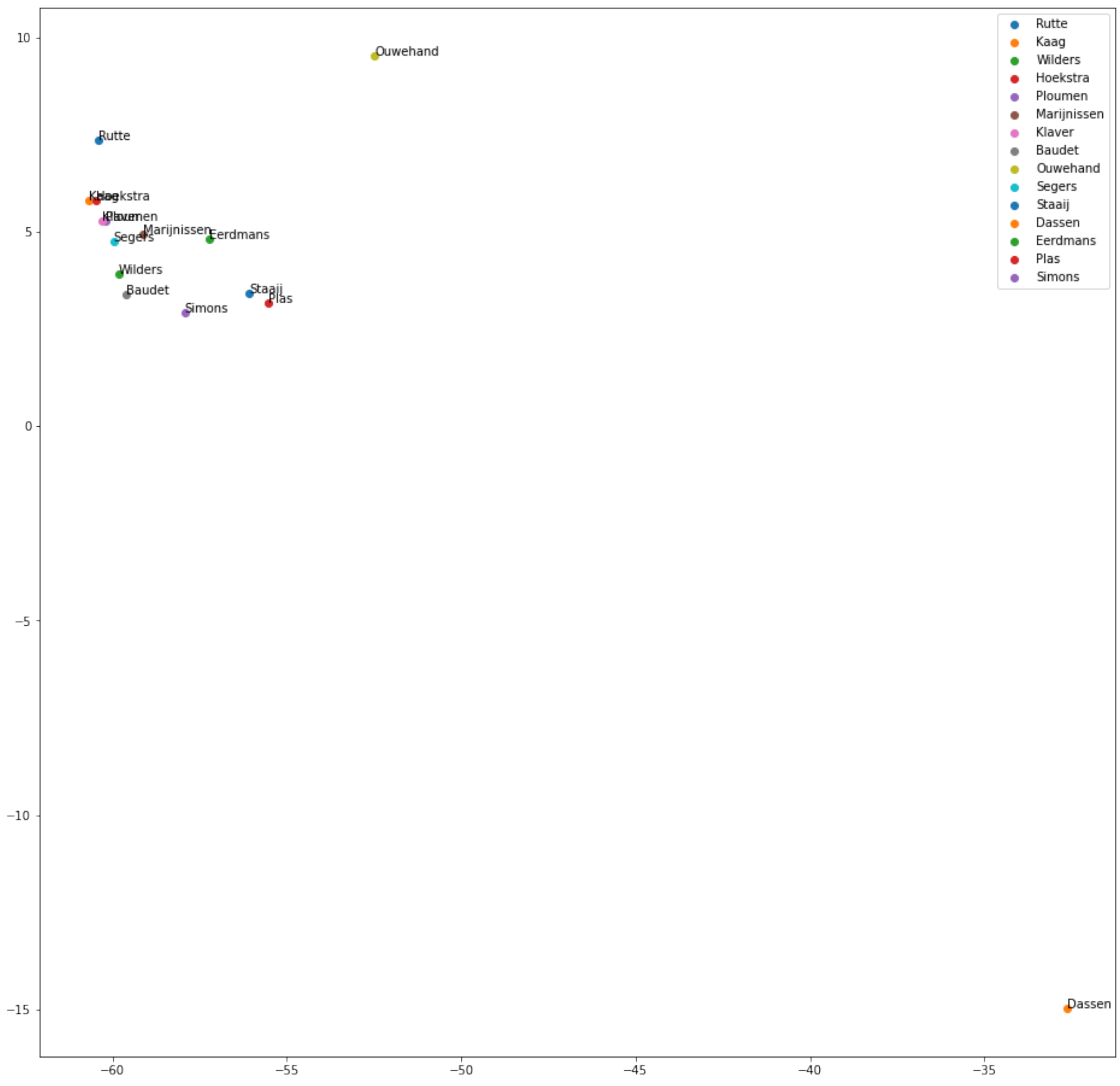


Figure 33: Word2Vec + tSNE: Politicians in Trouw

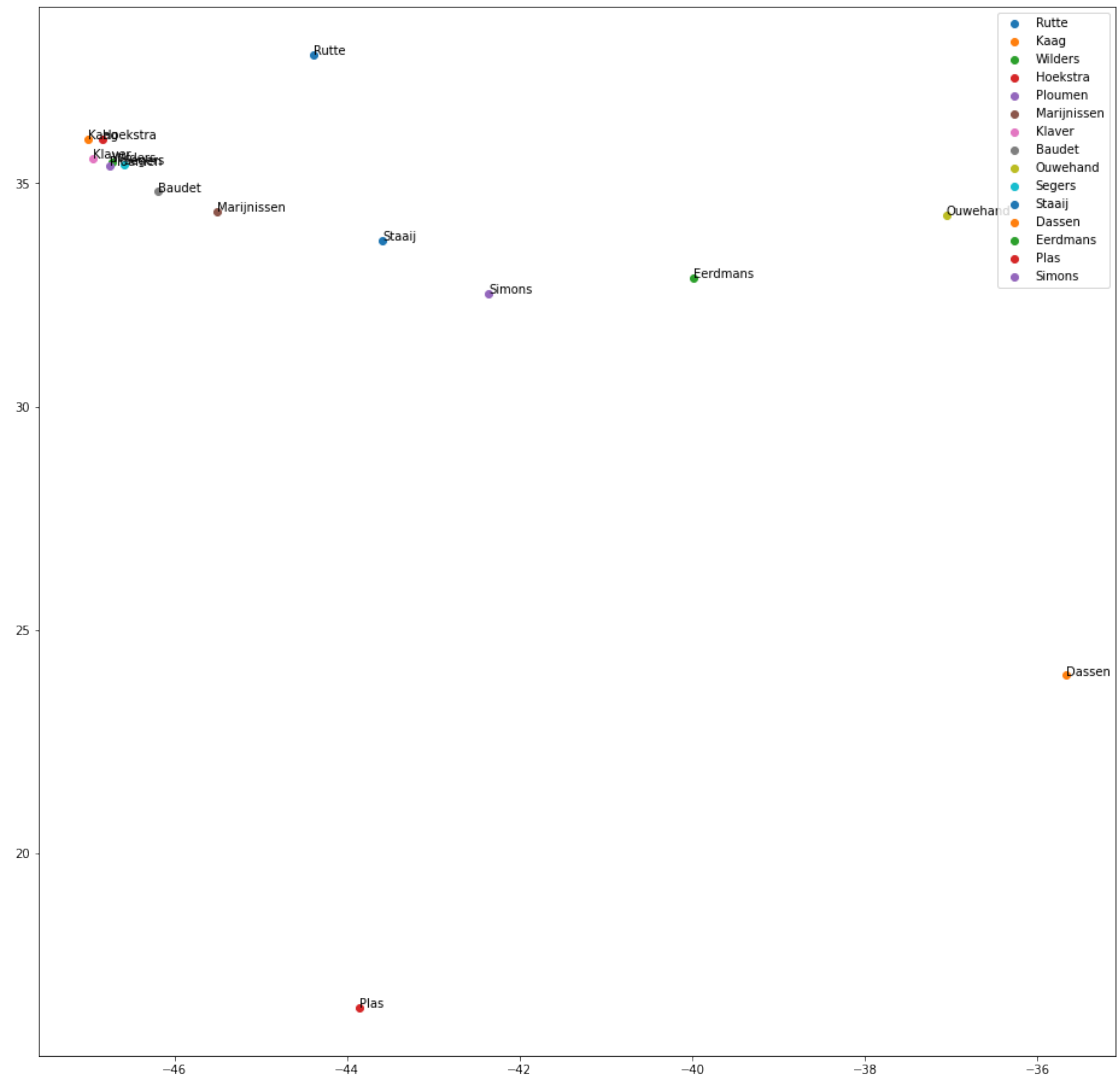


Figure 34: Word2Vec + tSNE: Politicians in Het Parool

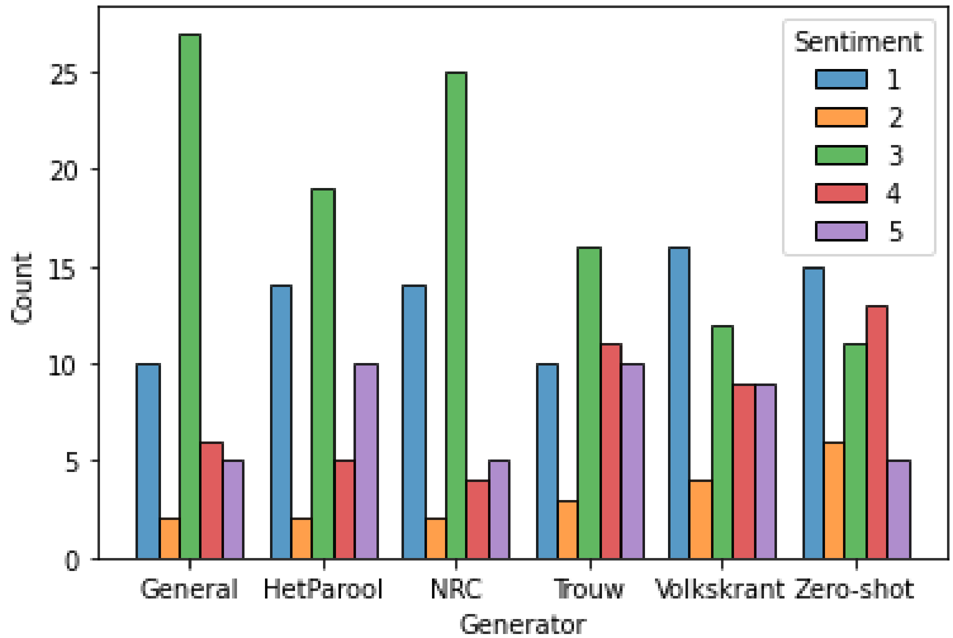


Figure 35: Sentiment analysis generative models VVD

## 9.2 Sentiment Divergence

Further visualisations as mentioned in 5. Furthermore, considering the parties, in figure 35, an interesting divergence is seen as Volkskrant generates a majority of text with a negative sentiment and Het Parool and NRC lean towards negative as well. From the newspapers, Trouw is the most positive towards VVD. Considering D66 in Figure 36, Trouw is most positive and most negative. The sampled text is overall neutral with a sentiment score of 3 having the majority for all sources. Considering CDA in Figure 37, Trouw is rather positive as the sentiment score of 4 is in the majority. Furthermore, the NRC leans negative with a significant amount of text scoring 1. Considering FVD in Figure 38, the overall sentiment is very negative. From the newspapers, Volkskrant relatively produces the most positive text.

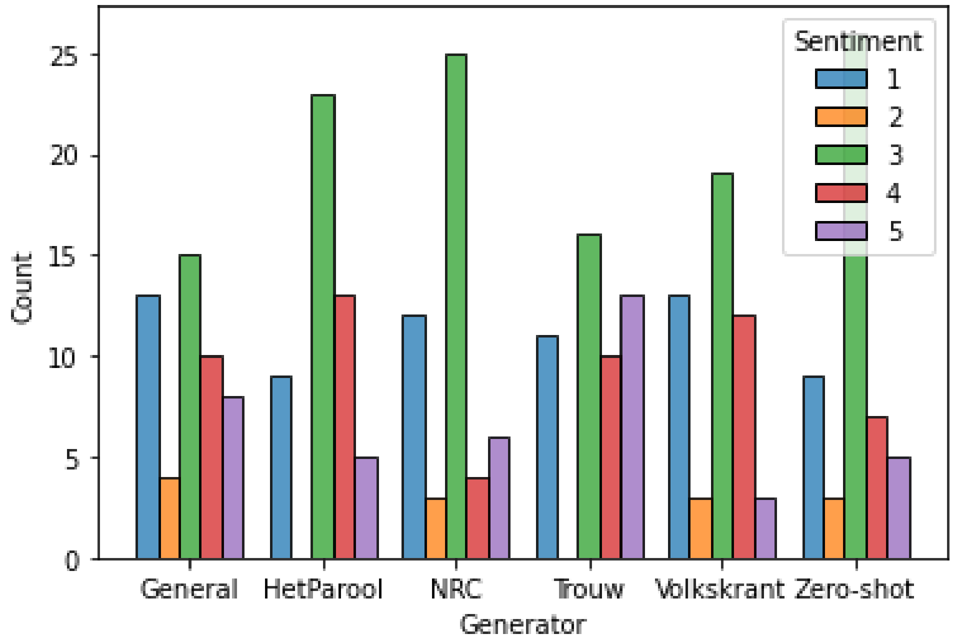


Figure 36: Sentiment analysis generative models D66

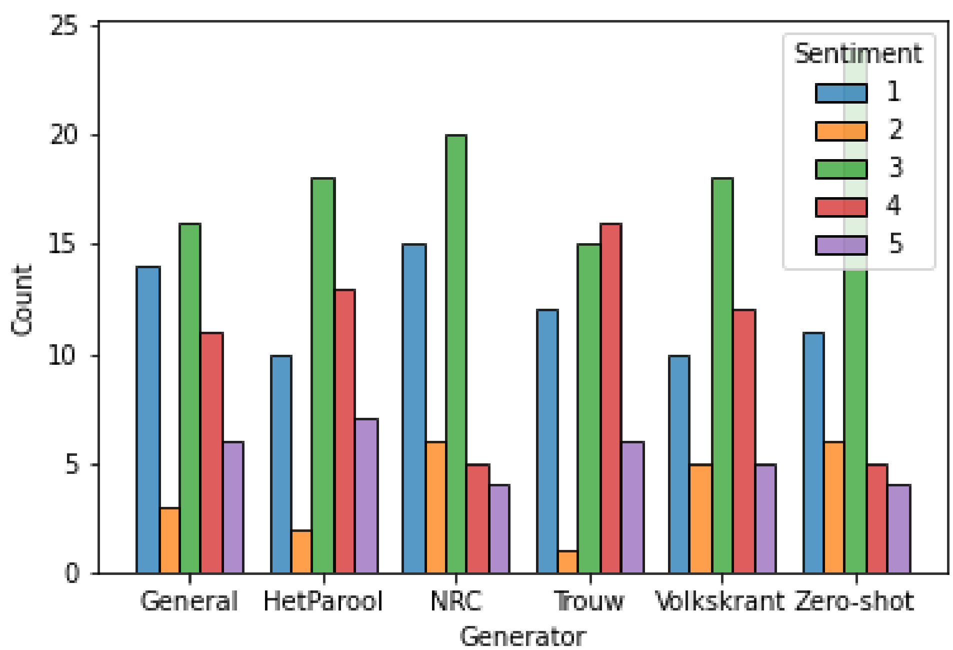


Figure 37: Sentiment analysis generative models CDA

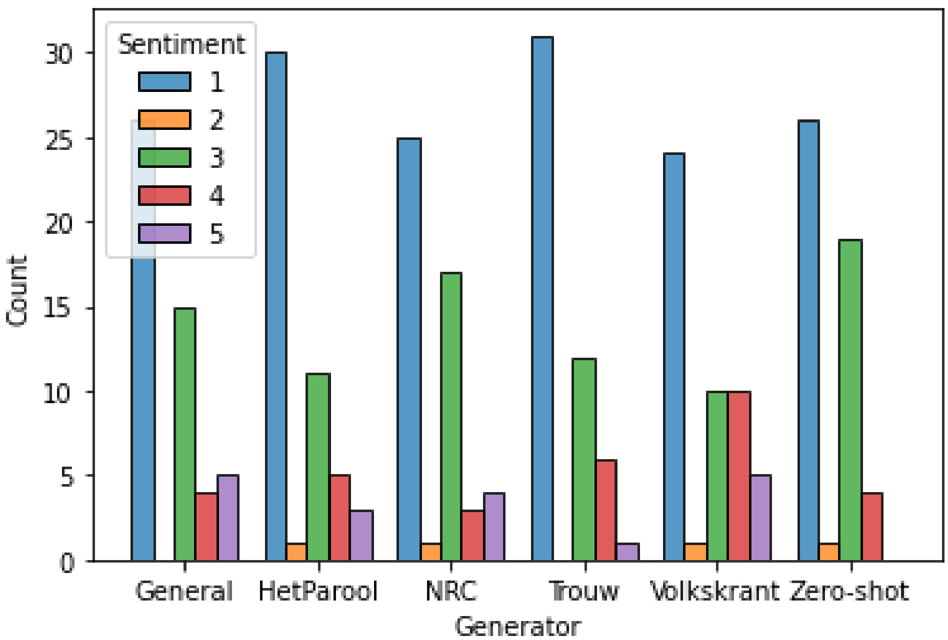


Figure 38: Sentiment analysis generative models FVD