Tracing Political Positioning of Dutch Newspapers

Christopher Congleton, Peter van der Putten(✉), and Suzan Verberne(✉)

Leiden University, Niels Bohrweg 1, 2333 Leiden, CA, The Netherlands
c.r.congleton@umail.leidenuniv.nl,
{p.w.h.van.der.putten,s.verberne}@liacs.leidenuniv.nl

Abstract. Newspapers write for a particular readership and from a certain ideological or political perspective. This paper applies various natural language processing methods to newspaper articles 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 by means of metrics, LDA topic modelling and word embeddings. Furthermore, article source discrimination is analysed by applying various classification models. Finally, 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.

Keywords: Political bias · Topic modelling · Newspaper agenda setting

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 if authors and media are transparent about their positioning [6], but ideological or political bias is an issue for analysis or news reporting articles [12]. Framing the debate and setting the political agenda offers media considerable influence depending on how critical the reader is in consuming content. Media outlets have 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

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

F. Spezzano et al. (Eds.): MISDOOM 2022, LNCS 13545, pp. 27–43, 2022.
https://doi.org/10.1007/978-3-031-18253-2_3

Table 1. Research questions

RQ1	*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 newspapers share a set of topics?*
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?*

newspaper's ideology or political perspective [12]. 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 [7], Denmark [10] and Korea [14] has quantified bias in seemingly politically neutral articles by means of modern computational techniques. Regarding selection bias work by Susanszky et al. [27] 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 vanilla GPT-2's bias in relation to occupation-gender ratio [17] as well as political bias [20]. Following up on GPT-2 with 1.5B parameters trained on 40 GB 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 570 GB [4,22].

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. This paper aims to bridge this gap by tracing political positioning of newspapers based on a large collections of their articles. Furthermore, in this paper we aim to retain the bias in the generative language model and analyse it in contrast to studies reducing bias in the model.

Work by De Vries et al. [30] 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 paper 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 paper seeks to answer the following key research question: to what extent is the ideological position of newspapers reflected in their writing? (for sub questions see Table 1).

The research questions are answered by applying computational techniques to a collection of 96,840 Dutch newspaper articles collected for the purpose of this paper. 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.

The contribution of this paper consists of introducing the application of specific 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 [16,23]. 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 paper is organized as follows. Section 2 reviews the background and related work. We then present methods, analysis and results for data collection (Sect. 3), coverage bias (Sect. 4, SQ1), discourse topic analysis (Sect. 5, SQ2-3), article source identification (Sect. 6, SQ4) and text generation (Sect. 7, SQ5), followed by a discussion (Sect. 8), limitations and future work (Sect. 9) and conclusion (Sect. 10).

2 Related Work

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

In the context of this paper, bias is the action of supporting or opposing a particular person or party in an unfair way through allowing opinion to influence judgment. It can manifest itself in various ways as illustrated in research by Eberl et al. [9] 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. [7] covering political bias in online newspaper articles uses occurrence metrics and sentiment analysis. Enevoldsen et al. [10] specifically use sentiment analysis to study tonality bias.

2.1 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 [28]. 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 [1,19] 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 [25].

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 using topic modelling or word embedding methods is not found, and thus leaves a void to be filled by this research paper.

2.2 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 [24] 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. [15] 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.

2.3 Text Generation

Modern natural language generation has many applications where text is generated from other forms of data or a language model [11]. In this paper, a transformer-based language model is used, specifically GPT-2 [22]. In this paper, 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. [21] which is based on the Transformer model [29]. In this paper 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.

3 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). We would have liked to have included more conservative, right-leaning or tabloid media.

3.1 Data Collection

The collection of articles for each newspaper is carried out following the same general sequence of steps[2] First, the website archive is crawled to index all the articles URLs in a specific time range (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. Capitalisation and punctuation are retained. Fourth, the data are saved as JSON dumps.

3.2 Results

In total 96840 articles have been collected, respectively for NRC (32043), Volkskrant (25702), Trouw (20944) and Parool (18151) as visualised in Fig. 1. To illustrate the size of these collections the number of words in each set is illustrated in Fig. 2. 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 3. To illustrate the relative size, the number of words for each source is visualised in Fig. 4.

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. Regarding 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.

[2] Sample code for data collection and analysis can be found at: https://github.com/Chris-Congleton/MSc-Thesis.

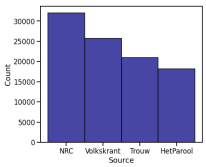

Fig. 1. Total number of articles

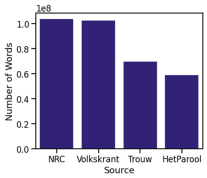

Fig. 2. Total number of words (in 100 Ms)

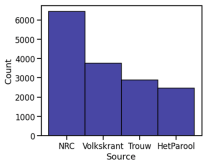

Fig. 3. Number of articles in political subset

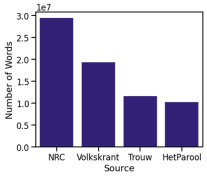

Fig. 4. Number of words in political subset (in 10 Ms)

4 Coverage Bias

Coverage bias (SQ1) is studied by evaluating various metrics based on the aggregation of party or politician mentions. The selection of party 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[3].

4.1 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 Eq. 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] The parties, party leaders and number of seats in the 2021 Tweede Kamer can be found at: https://www.kiesraad.nl/actueel/nieuws/2021/03/26/officiele-uitslag-tweede-kamerverkiezing-17-maart-2021.

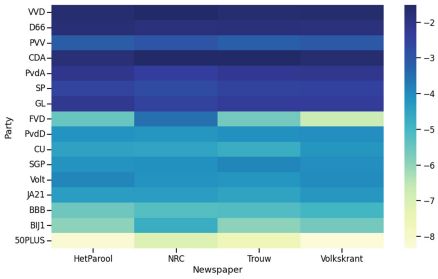

Fig. 5. Party-newspaper coverage

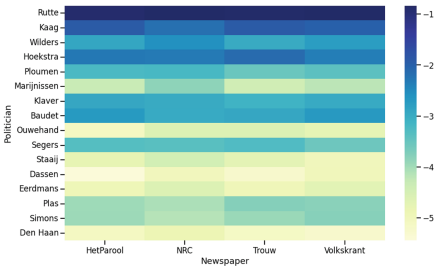

Fig. 6. Politician-newspaper coverage

4.2 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.

4.3 Experiments and Results

In order to illustrate the political coverage bias present in newspapers the results of the Log Normalised Mention Frequency of parties and politicians in 2021 are depicted in Figs. 5 and 6. 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 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 Fig. 6, 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.

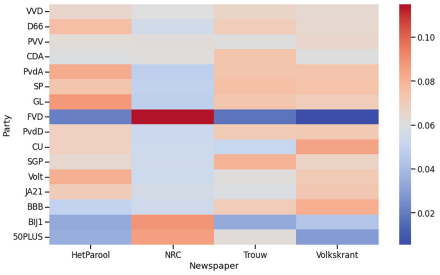

Fig. 7. Relative party-newspaper coverage

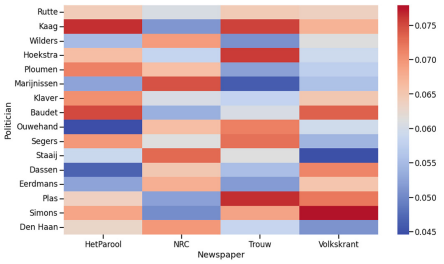

Fig. 8. Relative politician-newspaper coverage

Furthermore, the relatively high coverage of Segers (CU) could be attributed to the fact that his party was part of the government.

With an increased contrast, the Relative Normalised Mention Frequency, is depicted in Figs. 7 and 8. 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 (VVD and prime minister) who is covered fairly consistently over all sources, the contrasts in politician coverage are more prevalent. Kaag (D'66) for example, does relatively well in Het Parool and Trouw. Hoekstra (CDA) and van der Plas (BBB) 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 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. LDA is a generative probabilistic model of a corpus [13].

5.1 Topic Modelling

To prepare the text data specifically for the topic modelling 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, 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, the political subset as mentioned in Sect. 3 is used.

The LDA model is analysed using the pyLDAvis package. This depicts the Intertopic Distance Map and 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 [26]. 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 were 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.

5.2 Experiments and Results

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 modelling is analysed with regards to the topic modelling 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. 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.

5.3 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 spatial projection of the parties and politicians based on how newspapers write about them as an alternative to the Horseshoe model as introduced in Sect. 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.

Table 2. LDA topic modelling (translated into English)

	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	-

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. 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.

5.4 Experiments and Results

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

The general political discourse is visualised in Fig. 9. 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 Fig. 10. 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

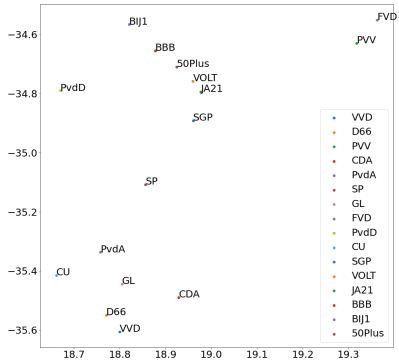

Fig. 9. Word2Vec + tSNE: parties in general

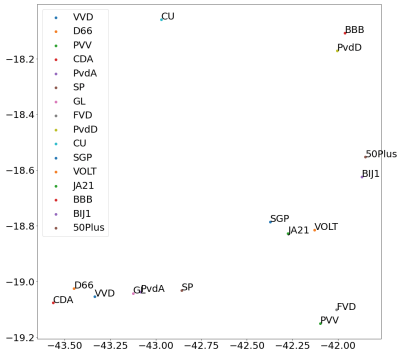

Fig. 10. Word2Vec + tSNE: parties in NRC

the middle right. Compared to the general political discourse visualised in Fig. 9, 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.

In conclusion, these results do not show a consistent shape of the parties in the embedding space. This is an indication against the presence of a generally shared political discourse.

6 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 labelled 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.

Experimentation is performed using various modelling techniques: Decision Tree, RBF Support Vector, XGBoost, KNeighbors, Gaussian Naive Bayes,

Table 3. Performance of source classification models

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

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 [8] 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.

6.1 Experiments and Results

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 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 two class combinations, the top ten positive and negative predictors are visualised. With respect to Parool-NRC, Fig. 11, it is logical to see 'amsterdam' as a strong positive feature and 'nrc' as a strong negative. With respect to Parool-Trouw in Fig. 12, it is interesting to see 'mark' a strong predictor for Parool in contrast with 'premier' for Trouw.

7 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

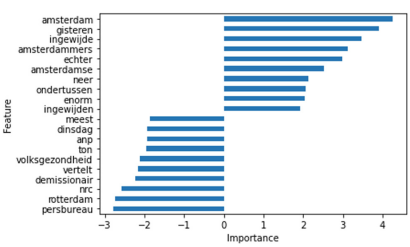

Fig. 11. Parool vs NRC

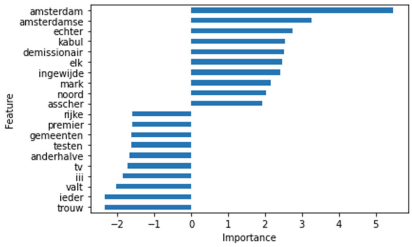

Fig. 12. Parool vs Trouw

style of an author or news medium. For sports articles automatic generation is already in use [18]. 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 [30]. This model recycles the original English GPT-2 model [22]. 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 (2.8 GB), newspaper articles (2.9 GB) from 2007, books (6.5 GB) and articles from various Dutch news websites (2.1 GB). 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.

This analysis using generative AI 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 prompt. Second, as the textual data are generated using 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.

7.1 Experiments and Results

The divergence in text generation of these models is compared through initiation of the different model versions with a neutral prompt. The model is implemented using the Huggingface's transformer package. The following subpackages are used: AutoTokenizer, TextDataset, DataCollatorForLanguageModeling, Trainer, TrainingArguments and AutoModelWithLMHead. The model is

Fig. 13. General Wordcloud

Fig. 14. NRC Wordcloud

fine-tuned using the full collection of articles as described in Sect. 3. 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.

To analyse the generative models with a neutral prompt, sampling is performed based on: "X houdt een toespraak" ("X gives a speech"). A set of 1000 samples is produced with maximal sequence length set to 30. From these texts, word clouds are produced where the most prominent terms are displayed scaled to their occurrence using the Wordcloud python package, with prompt words removed. One of the generated samples is (translated from Dutch): "X is giving a speech on the developments around the corona virus in his capital, The Hague. On social media, he has criticised politicians who do not get themselves investigated together whether they can use corona vaccinations to prevent that".

As can be seen from this example, the specific capital mentioned is not correct. Still, the usage of terms and coverage of topics can provide insight.The word cloud based on the general text generator model, Fig. 13, 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. These results show a significant divergence in topics resulting from a neutral prompt. We have carried out a range of other generative experiments that have been omitted here for brevity, for full details see [5] (Fig. 14).

8 Discussion

Concerning SQ1, on political bias, some results stand out. A low coverage in relation to the 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 political discourse, national politics is the most prominent topic across all newspapers. However, other topics differ significantly.

Regarding the vector space representation of 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. Taking into account topic modelling and vector representation analysis, the newspaper-specific political discourse differs considerably from the general political discourse.

Concerning SQ4, on the identification of a newspaper given an article, this task was very manageable for the advanced RobBERT model. For simpler methods only moderate accuracy 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. 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 consider that this type of research is more speculative as it based on generated, synthetic data (see [5] for more results).

9 Research Limitations and Future Work

The analysis in this paper 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 paper 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.

10 Conclusion

In this paper, we have analysed the extent to which the ideology of a newspaper 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 generative models trained on articles. The results showed several indications of political tendencies: disproportionate coverage of politicians and parties, limited overlap of political discourse, classifiable article source and divergence of generated text. Even though it is generally known that newspapers write from an ideological and political point of view, 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 media, and as in our new analysis, use a multitude of tools to analyse 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. Chin. Polit. Sci. Rev. **3**(4), 464–494 (2018)
2. Abuhamad, M., Rhim, J., AbuHmed, T., Ullah, S., Kang, S., Nyang, D.: Code authorship identification using convolutional neural networks. Futur. Gener. Comput. Syst. **95**, 104–115 (2019)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Congleton, C.: Tracing Political Positioning. Master's thesis, Leiden University, July 2022
6. Coppock, A., et al.: The long-lasting effects of newspaper op-eds on public opinion. Q. J. Polit. Sci. **13**(1), 59–87 (2018)
7. 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)
8. 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, November 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.292
9. 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. Int. J. Press/Polit. **22**(1), 111–132 (2017)
10. Enevoldsen, K.C., Hansen, L.: Analysing political biases in Danish newspapers using sentiment analysis. J. Lang. Works-Sprogvidenskabeligt Studentertidsskrift **2**(2), 87–98 (2017)
11. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018)
12. Hassell, H.J., Miles, M.R., Reuning, K.: Does the ideology of the newsroom affect the provision of media slant? Polit. Commun. **39**(2), 184–201 (2022)
13. Jelodar, H., et al.: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools App. **78**(11), 15169–15211 (2019)
14. Kang, H., Yang, J.: Quantifying perceived political bias of newspapers through a document classification technique. J. Quant. Linguist. **29**, 1–24 (2020)
15. Kats, T., van der Putten, P., Schelling, J.: Distinguishing commercial from editorial content in news. In: Preproceedings 33rd Benelux Conference on Artificial Intelligence and the 30th Belgian Dutch Conference on Machine Learning (BNAIC/BENELEARN 2021), Luxembourg, 10–12 November 2021 (2021)
16. Kawintiranon, K., Singh, L.: PoliBERTweet: a pre-trained language model for analyzing political content on twitter. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) (2022)
17. Kirk, H.R., et al.: Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. Adv. Neural. Inf. Process. Syst. **34**, 2611–2624 (2021)
18. Kunert, J.: Automation in sports reporting: strategies of data providers, software providers, and media outlets. Med. Commun. **8**(3), 5–15 (2020)

19. 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, pp. 447–456. UAI 2016, AUAI Press, Arlington, Virginia, USA (2016)
20. Liu, R., Jia, C., Wei, J., Xu, G., Vosoughi, S.: Quantifying and alleviating political bias in language models. Artif. Intell. **304**, 103654 (2022)
21. Radford, A., et al.: Improving language understanding by generative pre-training (2018)
22. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
23. Ressmeyer, R., Masling, S., Liao, M.: "Deep faking" political twitter using transfer learning and GPT-2 (2019)
24. 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. Futur. Internet J. **13**(1), 3 (2020)
25. 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, October 2020
26. 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)
27. Susánszky, P., Kopper, Á., Zsigó, F.T.: Media framing of political protests-reporting bias and the discrediting of political activism. Post-Soviet Affairs. 1–17 (2022)
28. Tangian, A.: Visualizing the political spectrum of Germany by contiguously ordering the party policy profiles. Data Anal. App 2. Util. Results Eur. Top. **3**, 193–208 (2019)
29. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
30. 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, pp. 836–864 (2020). https://arxiv.org/abs/2012.05628