The MaML datasets: Political actors, topics and tone in 15 newspapers from Norway, Denmark, Belgium, the Netherlands and United Kingdom

Gunnar Thesen
University of Stavanger, Norway
gunnar.thesen@uis.no

Erik De Vries
University of Stavanger, Norway
erik.devries@uis.no

January 2024

For using any of the described data, in parts or in full, or for quoting this note, please cite the resource as:

Thesen, Gunnar and Erik De Vries (2024) 'The MaML datasets: Political actors, topics and tone in 15 newspapers from Norway, Denmark, Belgium, the Netherlands and United Kingdom', Harvard Dataverse, https://doi.org/10.7910/DVN/J6F4PP

Content

1. Introduction 2
2. Selection of news sources and corpus construction 3
3. Coding the news appearances of political actors 6
4. Coding the topic content in the news 7
5. Coding the sentiment in the news 10
6. Overview of files in dataverse 11
7. Funding and acknowledgements 12
8. References 13

1. Introduction

This note documents the data collection for a comparative project on news and party support entitled "Media as the missing link" (MaML). The data comprises a corpus of newspaper articles from Norway, Denmark, Netherlands, Belgium and UK, covering the period from 2000 until 2018/19. Sources were selected based on de Vreese et al (2017): Comparing Political Journalism. The corpus covers the full content of three newspapers from each country: 1 mass-market, 1 left-leaning and 1 right-leaning upmarket.

Section 2 elaborates on the choice of news sources and the processes involved in the preparation of the corpus. Section 3 to 5 subsequently explain the coding of the three sets of variables that has been added to the corpus through a combination of automated and manual coding processes:

- 3, the presence of political actors in the news
- 4, the topic content in the news
- 5, the sentiment in the news

Data files and additional resources are available in our dataverse, see https://doi.org/10.7910/DVN/J6F4PP

Throughout the note, we refer to these resources in the relevant sections. Additionally, section 6 provides an overview of the files in the dataverse. Five country datasets are available for download, together with a brief documentation note explaining the datasets and the variables.

2. Selection of news sources and corpus construction

The selection of newspapers (see Table 2.1) was based on a recent comparative analysis of political journalism (De Vreese, Esser and Hopmann 2017) which included the leading left-leaning broadsheet, the leading right-leaning broadsheet and one mass-market newspaper in each country. However, we have made one change to the sources applied in De Vreese, Esser and Hopmann (2017). The left-leaning broadsheet Dagsavisen, which was part of the Norwegian sample in their study, was not an option due to low data quality and accessibility. One the one hand, this arguably less of a problem in Norway compared to the other countries, since there are no strong candidates for a left-leaning broadsheet with a sizeable and national circulation in Norway. Dagsavisen for instance, has a limited circulation which amounts to less than 10% of the leading right-leaning broadsheet Aftenposten. Furthermore, nearly all of its readers are located in the capital region (Høst 2019). In the end we chose to include the tabloid Dagbladet as a replacement, meaning that we capture a substantially larger share of the national news market. Dagbladet is clearly not a left-leaning newspaper in the same way as Dagsavisen, although it has occasionally issued editorial warnings against incumbent conservative coalitions in the period we cover (e.g., Pettersen 2009).

Table 2.1. News sources by country, political leaning and format.

Country	Left of center	Right of center	Mass-market	Period
Denmark	Politiken	Jyllandsposten	Ekstra Bladet	2000 to 2019
Belgium	De Morgen	De Standaard	Het Laatste Nieuws	2000 to 2019
Netherlands	De Volkskrant	NRC Handelsblad	De Telegraaf	2000 to 2018
UK	The Guardian	Daily Telegraph	The Sun	2000 to 2018
Norway	*Dagbladet**	Aftenposten	VG	2000 to June 2019

For each of these 15 sources, the corpus contains all news published in the period 2000 until 2018/19, summing to a total of over 7 million news articles. Note that this is the final count of articles *after* pre-processing the corpus. The pre-processing contains several important steps. First, items with a word count below 30 have been removed from the corpus, because they nearly always refer to longer articles on other pages in the newspaper. The number of items removed through this procedure was approximately 641.000.

Second, the original newspaper data contained a substantial number of duplicate articles (within the same day). Duplicates were not randomly distributed across time and sources. Therefore, to increase the validity of comparisons, duplicate entries have been removed based on the cosine similarity of article pairs published on the same day by the same newspaper. When article pairs have a similarity of 0.85 or above, one of the articles is (at random) removed from the dataset. These similarity scores have been based on the first 300 words of each article. This cutoff was chosen to reduce computational complexity, and to make comparisons between short and long articles more equal. For instance, we encountered cases where one article contained more words than the other, but where the articles were still clearly duplicates of each other. Had we used the full article text of both documents, their

cosine similarity might not have been high enough to detect them as duplicates. The number of articles removed through this procedure was 964.311.

Next, the remaining articles went through Natural Language Processing (NLP) using the R package UDPipe (Straka & Straková, 2017) and version 2.3 of the Danish DDT, Dutch Alpino, Norwegian Bokmål and English EWT Universal Dependencies Models (Nivre et al., 2018). The goal of this procedure is to remove the inherent complexity of natural language by reducing all words in an article to their dictionary lemma. In addition, the process of NLP produces Universal Part-Of-Speech (UPOS) tags for each lemma, indicating the function of the word in a sentence. In this way, words that are written the same way but carry different meanings can be disambiguated.[1]

Finally, before analyzing and coding the corpus, we also removed news stories that deal mainly with sports and entertainment. To classify these ("irrelevant") articles, around 12,000 news articles have been hand-coded in English, and between 6,000 and 7,000 in Danish, Dutch and Norwegian. The reason for the difference between English and the other languages is because similar classification performance for all countries needs to be obtained, and this required more data in English than the other languages. Research assistants have classified these articles based on the categories "Culture/art events and entertainment", "Sporting events and athletes" and "Miscellaneous". If articles fall into any of these three categories, they are considered irrelevant, if not, they are relevant. The miscellaneous category contains all articles that cannot be classified in any of the other categories in the adapted Comparative Policy Agendas codebook (see /Topics/Codebook for manual topic coding of MaML news.pdf). Prior to the coding, the RAs went through several rounds of training. The intercoder reliability scores for this coding reached a level of 0.81 to 0.86, see table 2.2 below.

*Table 2.2. Intercoder reliability human coding of irrelevant articles**

Country	Final round	Min	Max	No. of coders	No. of training rounds
Belgium	0.84	0.77	0.87	7	5
Denmark	0.86	0.81	0.88	4	3
Netherlands	0.81	0.71	0.89	7	3
UK	0.86	0.84	0.89	6	3

* Intercoder reliability measured with Krippendorff's alpha. N=200 for each round of coding. Alpha in table refers to all RAs and one project member. Min refers to lowest score among all coders in final round, max refers to highest score. For Norway, irrelevant articles were not filtered out in advance but instead during the topic coding process.

The hand-coded articles are then used as input for a Naive Bayes classifier. The input features for this model are the tf-idf weighted lemmas and UPOS tags generated in the NLP procedure described in the paper. The "format" of each word/feature in an article becomes lemma_UPOS. For getting the best-performing model for each country, a 3 by 5 nested cross-validation procedure is used, with the 3 outer folds being used for performance estimation of the final model, and the 5 inner folds of each outer fold being used for parameter optimization. In this case, parameter optimization consists of only a single parameter, for feature selection. Features are selected based on the chi2 measure to determine

[1] Eg. book as a noun and book as a verb.

which features are most and least strongly associated with the “irrelevant” topic. Using the absolute chi2 values, the top x-th percentile of features are kept to construct a model. Through the nested cross-validation procedure described above, the optimum cutoff values for feature selection are determined as follows: 0.99 (BE), 0.995 (DK), 0.996 (NL), 0.994 (NO), 0.994 (UK). Using these parameters, the final models achieve a precision of between 0.87 (DK) and 0.94 (UK). Precision is used as optimization measure to avoid as much as possible that relevant articles are classified as irrelevant, allowing for some relevant articles to remain in the relevant articles category. More details on the process of removing irrelevant article can be found in De Vries (2022).

3. Coding the news appearances of political actors

The news appearances of political actors are captured by running queries in the corpus. Political actors are limited to political parties and individual politicians serving as either MPs, ministers, or party leaders.

Mentions of political parties are collected using case-sensitive queries on either the full party name, or the most commonly used party abbreviations. When necessary, special characters like opening and closing brackets for the abbreviations (con) and (lab) in the UK, are also taken into account. In Norway and Denmark, several of the major political parties have single letter abbreviations. In these specific cases, regular expression filters are used to filter out common mistakes, like V (the abbreviation for the left-wing party Venstre) as a roman number 5 in the names of monarchs.

Queries for individual politicians (ministers, party leaders and MPs), are constructed by looking for the combination of the (first) given name and surname within 5 words of each other. A larger distance between the two would result in too many false positives, and a smaller distance in too many false negatives. The queries are also limited to articles published during the time the politician was in office. For ministers the queries include their formal title as an alternative for their given name (e.g. both Secretary Johnson and Boris Johnson are valid hits).

A list of the parties and politicians that were queried is available in our dataverse (/Actors/maml_actors.xlsx). This list also contains the regular expression filters used for dealing with common mistakes.

4. Coding the topic content in the news

The content of each article in the corpus has been coded according to an abbreviated and adjusted version of the classification scheme applied by the Comparative Policy Agendas project (www.comparativeagendas.net). The codebook and coding instructions are available in our dataverse, see /Topics/Codebook for manual topic coding of MaML news.pdf. The procedure involves handcoding a random sample for each language containing between 35 to 45 000 news articles. These samples are subsequently used to train an automated classifier. Before handcoding the sample, RAs were trained in several rounds. The level of intercoder reliability in the final test was between 0.71 and 0.81, see table 4.1 below.

*Table 4.1. Intercoder reliability human coding of article topic**

Country	Final round	Min	Max	No. of coders	No. of training rounds
Norway	0.81	0.79	0.86	3	4
Belgium	0.71	0.68	0.74	4	6
Denmark	0.79	0.76	0.81	4	7
Netherlands	0.72	0.66	0.77	7	9
UK	0.81	0.81	0.82	3	6

* Intercoder reliability measured with Krippendorff's alpha on the major topic level. N=200 for each round of coding. Alpha in table refers to all RAs and one project member. Min refers to lowest score among all coders in final round, max refers to highest score.

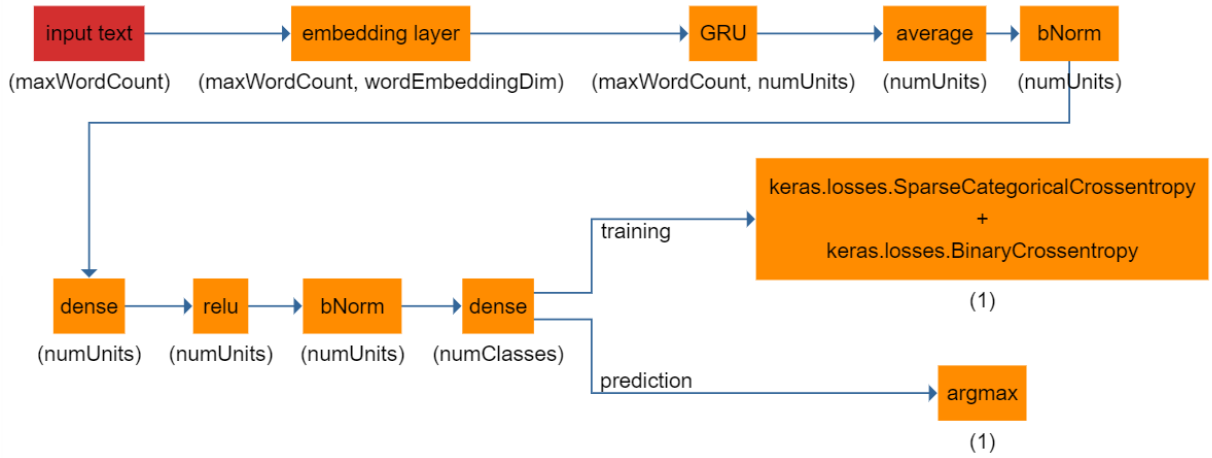
Prior to training the automated classifier (described below), a few changes were made to the handcoded data in order to reach a satisfying level of performance. It is important to be aware of these adjustments when using the MaML-data in combination with other CAP-coded data:

- All subtopic labels were recoded to their major topic label, eg. 101 to 1, 1302 to 13 etc.
- Major topic 18 (Foreign trade) and 21 (Public land, spatial planning and resource management) occurred with a very low frequency in the corpus, causing problems with their classification. As a consequence, these two categories were merged with the respective major topics which they are most similar to. Hence 18 was relabelled 19, while 21 was relabelled 7. This means that the automated classification of topic 19 indicates articles on foreign trade as well as foreign affairs, development aid and international economy. While the automated classification of topic 7 indicates public land etc in addition to environment.

Finally, note that topics 91, 92 and 93 of the codebook are dealt with through the classification of 'irrelevant' articles (see section 2). These categories are therefore not applied in the automated topic coding.

To classify the topic of an article, and articles that are exclusively about foreign news topics (domestic vs non-domestic), we use the deep learning framework Keras (Srinivasa-Desikan 2018). The procedure follows the data flow illustrated in Figure 4.1 below.

Figure 4.1. The data flow of the topic classification procedure.



The input text vocabulary is limited to the 10.000 words most commonly found in the train data, and the input text is limited to the first 800 words (maxWordCount). From this input, a word embedding model is generated, using 512 dimensions (tuned). The word embedded representation of the articles is then used as input for a recurrent neural network as described in Cho et al. (2014). After passing through some intermediate neural network layers, the dense layer produces a vector of predicted probabilities for each document, for both the topic and foreign news variables. During training, these are parsed through a sparse categorical crossentropy loss function for the topic labeling, and through a binary crossentropy loss function for the foreign news labeling. Both are added together, and the foreign news loss is multiplied by a factor of 7 (tuned). For labeling, the label with the highest probability is selected.

Optimization of this model has been conducted using 90% of the manually labeled data as training data, and 10% as testing/validation data. Once the optimum model parameters have been determined, these are used in a 10-fold cross-validation procedure to determine the final performance of the model. While it is true that this 10-fold CV includes the 10% test data used to determine the optimal parameters, the stability in performance between the individual folds leads us to the conclusion that the performance estimates provided by the 10-fold CV procedure are reliable.

Using this model, the resulting topic classifications achieve a weighted F1 (the harmonic mean of precision and recall) of between 0.64 (NL/BE) and 0.69 (UK). Given that the model is predicting 22 topic classes, and the considerable complexity of news articles that usually touch upon several topics, this is an acceptable performance. Table 4.2 displays the results per topic category and country. Wojcieszak et al (2021) report exceptionally high performance (accuracy up 0.78) when applying the CAP codebook to tweets, which are shorter than news articles and easier to classify. We are not aware of machine learning classifiers performing substantially better than ours when categorizing traditional news articles based on a multiclass scheme with a high number of topics that are also highly unbalanced (see Burscher et al. 2015). Sebők, M., & Kacsuk recently succeeded in assigning CAP-labels with a precision of over 80 % for most topics, but this approach leaves over 40 % of their corpus unlabeled.

Table 4.2. Classifier performance (F1), across languages and topics.

Code	Label	BEL/NLD	DNK	NOR	GBR
1	Macroeconomics	0.52	0.49	0.51	0.57
2	Civil rights and liberties	0.32	0.40	0.43	0.46
3	Health	0.75	0.76	0.76	0.80
4	Agriculture, fisheries and food	0.55	0.64	0.64	0.51
5	Labour	0.58	0.54	0.59	0.50
6	Education	0.76	0.76	0.76	0.81
7	Environment	0.53	0.61	0.54	0.62
8	Energy	0.59	0.65	0.66	0.70
9	*Refugees and immigration*	*0.65*	*0.63*	*0.66*	*0.66*
10	Transport	0.70	0.73	0.72	0.67
12	Crime and justice	0.77	0.75	0.82	0.83
13	Social welfare and social affairs	0.47	0.55	0.55	0.52
14	Housing and urban/rural development	0.50	0.62	0.66	0.71
15	Commerce, banking and consumer issues	0.53	0.70	0.62	0.68
16	Defense and security	0.74	0.71	0.74	0.76
17	Research, technology, IT and mass media	0.56	0.57	0.63	0.67
19	Foreign affairs, trade, international economy	0.62	0.55	0.55	0.55
20	Public sector and politics in general	0.74	0.64	0.69	0.70
23	Culture, art	0.44	0.44	0.69	0.52
24	Sports	0.60	0.71	0.74	0.67
25	Natural disasters, fires, preparedness	0.61	0.53	0.63	0.63
26	Religion and churches	0.58	0.61	0.71	0.56
	Average performance (weighted F1)	*0.64*	*0.64*	*0.67*	*0.69*

5. Coding the sentiment in the news

To measure the tone or sentiment in the news we rely on a method involving word embedding models similar to recent applications in political communication and political science (Rheault and Cochrane 2020; Rudkowsky et al. 2018). The method was proposed by Rheault et al. (2016) and further developed by Erik de Vries for this project. A detailed account of the approach can be found in De Vries (2022). Note that code and replication material can be found at GitHub - vriezer/sentiment: Replication materials

The process starts with a small and context-independent "seed dictionary" containing one hundred unambiguous negative and positive words. The seed dictionaries for the different languages are available in the dataverse. This is used to build an extensive sentiment dictionary adapted to the news corpus. We do this with the help of machine-learning models that extract meaning from a text by estimating a multi-dimensional vector space in which each word is positioned. Essentially, this approach finds the words that appear in proximity to our words of interest from the seed dictionary, based on the assumption that neighboring words will share an association to the latent semantic meaning of a piece of text (Mikolov et al. 2013).

The result is a longer dictionary (also available in the dataverse) with sentiment values for each word: higher positive values indicate closer proximity to the positive seed words, and higher negative values indicate closer proximity to the negative seed words. Based on the expanded dictionary, we then calculate how positive or negative each sentence and article in our corpus is.

The procedure is validated by having trained research assistants code a random sample of sentences. The instructions for this coding can be found in our dataverse, see /Tone/Codebook for tone coding.pdf. The intercoder reliability ranged from 0.71 (UK) through 0.75 (Denmark) and 0.79 (Norway) to 0.84 (Dutch). Comparing the human-coded sentiment to the sentiment based on the word embedding dictionary, our F1-scores range from 0.61 to 0.64.

De Vries (2022) provides more details on the validation, in addition to examples of applications that indicate good predictive validity. For instance, the well-established negativity bias of political news can be reproduced with our data. Furthermore, in each country, the tabloid is more negative than the broadsheets. Summarizing, the performance is on a par with the best performing non-manual sentiment model (based on non-human coding) tested in a recent study by Van Atteveldt, van der Velden, and Boukes (2021, see overview of results page 128).

6. Overview of files in dataverse

Table 6.1. File location, names and content.

Location	Name	Content
Root folder	Thesen and De Vries MaML dataverse documentation 2024.pdf	The present document
/Actors/	maml_actors.xlsx	List of all parties and individual politicians queried in the corpus
/Topics/	Codebook for manual topic coding of MaML news.pdf	The adjusted CAP classification scheme, including coding instructions
	Intercoder_Reliability_Irrelevant_Articles.xlsx	Intercoder reliability scores for human coding of irrelevant articles
	Intercoder_Reliability_Major_topics.xlsx	Intercoder reliability scores for human coding of major topics
	About the topic classifier.pdf	Short description of the deep learning framework applied by Markus Fjellheim for the topic classification
	performance_topic_classifier.csv	Performance scores for the topic classifier
/Tone/	Codebook for tone coding.pdf	Coding instructions for human coding of tone in the news
	Intercoder_Reliability_Tone.xlsx	Intercoder reliability scores for human coding of tone
	validation_results_sentiment_*country*.Rds	Four rds-files containing the results from validating the automated sentiment coding against human coded sentiment.
/Tone/ Dictionaries/	seed_dict_*country*.csv	Four seed dictionaries (benl=Belgium and the Netherlands, no=Norway, dk=Denmark, uk=United Kingdom) used as a starting point to create full sentiment dictionaries.
	full-lexicon-*country*.txt	Four full sentiment dictionaries created through the word embedding models
/Data/	article_level_*country *.dta	Five country datasets in Stata-format, containing information on the article X actor level
	Documentation for article level MaML datasets.pdf	Documentation of the article X actor level datasets and variables

7. Funding and acknowledgements

The MaML-project was funded by the University Fund for Rogaland, Stavanger, Norway.

We are indebted to the rest of our project group who has coordinated data collection efforts outside Norway: Christoffer Green-Pedersen and Peter Bjerre Mortensen in Denmark; Annelien Van Remoortere, Rens Vliegenthart and Stefaan Walgrave in Belgium and the Netherlands; Will Jennings in the UK.

A special thanks to data analyst Markus Fjellheim who is responsible for the topic classifier described in section 4.

Finally, we are grateful for the excellent work of our 25+ research assistants who have coded news across the five countries.

8. References

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts?. The ANNALS of the American Academy of Political and Social Science, 659(1), 122-131.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

De Vreese, C., Esser, F., & Hopmann, D. N. (eds). (2017). Comparing political journalism. London: Routledge.

De Vries, E. (2022). The Sentiment is in the Details: A Language-Agnostic Approach to Sentence-Level Sentiment Analysis in News Media. *Computational Communication Research*. doi:10.31235/osf.io/8y3jq.

Høst, S. (2019). Papiraviser og betalte nettaviser 2018. Statistikk og kommentarer. Report no. 90/2019, Volda University College.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21.

Pettersen, Ø. B. (2009). Pressen og partiene - Partitilhørighet i 2005. En analyse av Aftenposten, Dagsavisen, Dagbladets og Dagens Næringslivs leder- og kommentarartikler under valgkampen i 2005. Master Thesis, Department of Media and Communication, University of Oslo.

Sebők, M., & Kacsuk, Z. (2021). The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. *Political Analysis*, 29(2), 236-249.

Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis* 28 (1): 112–33.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PloS ONE* 11 (12): e0168843.

Rudkowsky, E., Haselmayer, H., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures* 12 (2-3): 140–57.

Van Atteveldt, W., van der Velden, M., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures* 15 (2): 121–40.

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2021). Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter.

Zeman, D., Nivre, J., & Abrams, M. (2019). "Universal Dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (UFAL)." Faculty of Mathematics and Physics. Charles University.