Amsterdam
University
Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 424-462
HTTPS://DOI.ORG/10.5117/CCR2022.2.003.VRIE

The Sentiment is in the Details

A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media

Erik de Vries
Department of Media and Social Sciences, University of Stavanger
erikdevries@uis.no

Abstract
Determining the sentiment in the individual sentences of a newspaper article in an automated fashion is a major challenge. Manually created sentiment dictionaries often fail to meet the required standards. And while computer-generated dictionaries show promise, they are often limited by the availability of suitable linguistic resources. I propose and test a novel, language-agnostic and resource-efficient way of constructing sentiment dictionaries, based on word embedding models. The dictionaries are constructed and evaluated based on four corpora containing two decades of Danish, Dutch (Flanders and the Netherlands), English, and Norwegian newspaper articles, which are cleaned and parsed using Natural Language Processing. Concurrent validity is evaluated using a dataset of human-coded newspaper sentences, and compared to the performance of the Polyglot sentiment dictionaries. Predictive validity is tested through two long-standing hypotheses on the negativity bias in political news. Results show that both the concurrent validity and predictive validity is good. The dictionaries outperform their Polyglot counterparts, and are able to correctly detect a negativity bias, which is stronger for tabloids. The method is resource-efficient in terms of manual labor when compared to manually constructed dictionaries, and requires a limited amount of computational power.

Introduction

The availability of large amounts of text data, cheap computing power, and an abundance of analytical tools have all led to a rising interest in automated text analysis methods. Despite this interest, the development of automated methods is far from finished. Using automated methods to determine the positive or negative tone (sentiment) in the individual sentences of a newspaper article remains a major challenge. Manually created sentiment dictionaries (e.g. Soroka et al., 2015; Young & Soroka, 2012) are often used, but have been shown to perform badly when compared to the gold standard of human coding (Boukes et al., 2020). These dictionaries all consist of words that are manually selected based on human expertise, which is a time-consuming process. Computer-generated dictionaries are more time-efficient, and seem to work slightly better than manually constructed ones. Evaluation is however conducted based on simplified tasks such as distinguishing 'clearly positive [. . .] and clearly negative headlines' (Khoo & Johnkhan, 2018, p. 505). Even when using supervised machine learning to classify levels of negativity in parliamentary speeches, performance does not reach higher than .61 (F_1-score) (Rudkowsky et al., 2018). These examples show there is still a lot of room for improvement when it comes to analyzing sentiment in media texts.

Rheault et al. (2016) provide a method for such an improvement, with their application of a word embedding model in combination with a small dictionary of positive and negative 'seed words'. However, they do not apply their method to political news, and validate the performance of their method in a different domain (movie reviews) than the domain that is of substantive interest to them (political speeches). The main question is therefore if their method can be successfully applied to the domain of political news in multiple languages, and at the level of individual sentences instead of documents/newspaper articles.[1]

To answer this question, a dataset containing two decades of newspaper articles in four languages (Danish, Dutch, English, Norwegian) is used. These articles are taken from three different newspapers for each language (six for Dutch/Flemish). Word embedding models are constructed for each of these languages and used to generate sentiment dictionaries based on a small list of positive and negative seed words, replicating to a large extent the method described by Rheault et al. (2016). Concurrent validity is evaluated by comparing the dictionary-based classification to the gold standard of human-coded sentiment. Predictive validity is evaluated by testing two long-standing hypotheses concerning the negativity bias in political news. Finally, the performance of the method is compared to the performance of the Polyglot sentiment dictionaries (Chen & Skiena, 2014), which is one of the best performing dictionaries in the comparison by Boukes et al. (2020).

Sentiment Analysis

While there are many aspects to sentiment, it can generally be described as the 'attitude towards a particular target or topic' (Mohammad, 2016, p. 201). These attitudes are either evaluative or emotional in nature. Evaluative attitudes are based on a simple one-dimensional scale for judging whether something is 'good' or 'bad'. Emotion, on the other hand, is a multidimensional concept, making it much harder to measure using automated methods than one-dimensional evaluative attitudes. Even so, the evaluative aspect of attitudes still remains hard to analyze in an automated fashion.

For one, it is hard to determine the source and target of an evaluation. Semantic role labeling provides a possible solution for this issue, by aiming to extract source-subject-predicate structures from a sentence. One way in which this can be done is by using the syntactic dependencies between words (Shi et al.,2020), as van Atteveldt et al. (2017) successfully do. When the source, subject and predicate in a sentence are known, this information can be used to conduct stance detection. The goal of stance detection is to determine the evaluative stance of the source towards the subject, based on the predicate. However, the stance of a source cannot be directly derived from the words that are used. A negative statement might still contain a positive evaluation, such as in the sentence 'I am sad that Hillary lost this presidential race' (example from Aldayel & Magdy, 2021, p. 5). While this statement is negative, the implicit evaluation of the target (Hillary) by the source (I) is positive.

The above relates closely to the difference between evaluation and valence, or between 'good' and 'bad' versus 'positive' and 'negative'. The former depends on perspective, what is good for somebody can be bad for somebody else. The latter disregards perspective, and is solely based on the inherent positive or negative connotation present in a word or sentence. As this paper is concerned with the creation of sentiment dictionaries, the only aspect of sentiment that can be investigated is valence. And while valence can be combined with semantic role labeling, the election example from Aldayel & Magdy (2021) illustrates that even then it is not always possible to reliably determine stance. Thus, the operational definition of sentiment in this paper is limited to the sum of the positive and negative connotations of words in a sentence.

When using a dictionary for sentiment analysis, there are two main aspects to consider: 1) the construction and content of the dictionary, and 2) the specific domain to which it is going to be applied. Constructing a suitable sentiment dictionary for a specific task is complex, as words have different meanings in different domains. Thus, a sentiment dictionary needs to be domain-specific to some extent (Young & Soroka, 2012). Muddiman et al. (2019) show that manually constructed dictionaries work quite well when they are applied in a very specific domain. Boukes et al. (2020) however show that when using sentiment dictionaries in a more general way (i.e. applied to multiple newspapers, on a general (economy) topic), none of the tested dictionaries perform particularly well. This illustrates the tradeoff in dictionary construction between specificity and general applicability.

Manually constructing dictionaries is a time-consuming process because of this tradeoff, and automating the process of dictionary creation can save valuable time. An additional advantage of automation is that the dictionary can be based on the corpus to which it will be applied, ensuring at least some level of balance between applicability and domain-specificity. One way to construct a computer-generated dictionary is by expanding a short list of positive and negative seed words to a full dictionary through a word embedding (WE) model (Rheault et al., 2016). WE models (Mikolov et al., 2013; see Almeida & Xexéo, 2019 for an overview) make use of the distributional hypothesis, 'a word is known for the company it keeps' (Firth, 1957), to construct a multi-dimensional vector space in which each word is positioned based on its co-occurrences with other words. The assumption is that the dimensions in this vector space represent different latent aspects of meaning (Mikolov et al., 2013), implying that words that are close together in one or more of these dimensions share to a larger or smaller degree their semantic meaning with neighbouring words.

Considering that words with similar meaning occur closely together, it is possible to construct a sentiment dictionary using the words that are closest to the positive and negative words in the seed dictionary (Rheault et al., 2016). Of course, the words in the seed dictionary need to be positive or negative in all possible semantic contexts. Otherwise, the words most closely associated with an ambiguous seed word will also contain ambiguous meaning/sentiment, and thus cause a bias in the final dictionary.[2] Assuming bias is absent from the seed dictionary, the procedure described here allows for the creation of domain-specific sentiment dictionaries as defined by Young & Soroka (2012) from any large corpus of documents.

Assuming that a sentiment dictionary is domain-specific to the data it is applied to, the next question is to which unit of text it should be applied. Ideally, the units of text being analyzed contain only information needed to answer the research question, without any noise. This is of course not realistic, especially when considering that a single newspaper article generally contains references to multiple topics, events, and/or actors. Because each of these subjects are associated with their own sentiment, it makes sense to only analyze those parts of an article that actually relate to the subject of interest. This trend is also visible in media studies, shifting from documents as the unit of analysis (e.g. Bleich & van der Veen, 2018; Young & Soroka, 2012) to smaller units, such as sentences or headlines (Boukes et al., 2020; Khoo & Johnkhan, 2018; Rudkowsky et al., 2018; van Atteveldt et al., 2021). These smaller units are less likely to contain multiple subjects, and make it possible to determine only the sentiment in close proximity to the subject(s) of interest. If document-level metrics are required for further analyses, the scores of individual sentences can be aggregated into sentence groups, providing a sentiment score at the document level. As such, there are no theoretical downsides to analyzing sentiment at the sentence level. From a methodological perspective there is the downside of increasing the complexity of the analyses. But considering that more precise measures are generally preferred, this is an acceptable tradeoff. The question that remains is how well a WE sentiment dictionary applied to newspaper sentences works when compared to human coding.

> **RQ$_1$:** *How well do sentiment dictionaries based on word embeddings and seed dictionaries perform, when compared to human expert-coding?*

Another question is to what extent the proposed method outperforms other dictionary approaches. To test this, the performance of the WE dictionaries is compared to that of the Polyglot (Al-Rfou et al., 2013) sentiment dictionaries in Dutch, Danish, English and Norwegian (Chen & Skiena, 2014). These dictionaries, like the whole Polyglot project, are based on the most frequently used words in Wikipedia articles from a specific language. These words are used to construct a huge network of one- and bi-directional semantic links between words. By propagating from selected seed words (much as in the approach above), the final sentiment dictionary in each language is constructed. Boukes et al. (2020) show that the Polyglot dictionary is one of the best performing dictionaries for detecting positive and negative sentiment in Dutch economic news headlines.

> **RQ$_2$:** *How well do sentiment dictionaries based on word embeddings and seed dictionaries perform, when compared to the Polyglot sentiment dictionaries?*

Negativity bias

In addition to evaluating the concurrent validity of the WE dictionaries through the research questions formulated above, the concept of negativity bias is used to assess the predictive validity of the method. Negativity is a predominant feature of political news (see Lengauer et al., 2012 for an overview), which should be easily detected by the dictionaries. Furthermore, theories on hard versus soft news provide a clear expectation regarding the amount of negativity present in tabloid and broadsheet newspapers, as soft news is a hallmark of tabloid journalism (Otto et al., 2017). It is characterized by a focus on author opinion (Glogger, 2019) and emotion (Reinemann et al., 2012). Combined with the negativity bias in political news, it is therefore likely that tabloid newspapers are more negative in their coverage of political news than broadsheet newspapers.

> **H_1:** *Sentiment will be more negative than positive in political news coverage*

> **H_2:** *Sentiment will be more negative in tabloid newspapers than in broadsheet newspapers*

Data & Methods

In table 1, an overview is presented of the newspaper data used for each language, which runs from January 2000 until December 2019 unless otherwise noted. The division between left-wing, right-wing and tabloid newspapers is based on De Vreese et al. (2016).

In order to get a usable sentiment dictionary, the raw data is processed in five steps: 1) The raw newspaper articles are pre-processed, 2) the processed

Table 1. Newspaper sample

	Left-wing	Right-wing	Tabloid/Popular	Total articles[1]
Danish	Politiken	Jyllands-Posten	Ekstra Bladet	2.08
Dutch(NL)[2]	de Volkskrant	NRC Handelsblad	de Telegraaf	2.16
Dutch(BE)	de Morgen	de Standaard	Het Laatste Nieuws	2.18
English[2]	The Guardian	The Daily Telegraph[3]	The Sun	5.12
Norwegian	[4]	Aftenposten	VG / Dagbladet	2.28

Note:
[1] In millions
[2] Until December 2018
[3] From January 2001
[4] Due to lack of suitable data, Dagbladet as substitution

articles are used to create a (GloVe) word embedding model, 3) from the raw articles, sentences for validation are extracted and manually coded, 4) the word embedding model is combined with the seed word dictionary to create an expanded sentiment dictionary, 5) the validation data and expanded sentiment dictionary are combined to optimize the dictionary through feature selection and tuning the interpretation of the raw sentiment scores. These five steps are elaborately described below, followed by a short summary. The general steps involved in the dictionary expansion process (steps 4 and 5) are visualized in figure 1.

Pre-processing

In the first step, the raw newspaper articles are pre-processed for use in a word embedding model. The complexity of the articles is reduced by using Natural Language Processing (NLP) to convert inflected word forms to their dictionary lemmas. UPOS (Universal Part-Of-Speech) tags are extracted in this process to allow disambiguation between lemmas that are spelled the same way, but have different meanings (such as 'evening' or 'entrance' as either a noun or a verb). NLP also allows for more accurate identification of sentence borders. For example, periods in abbreviations and initials are not treated as sentence borders. NLP is conducted using the R package UDPipe (Straka & Straková, 2017), in combination with version 2.3 of the Danish DDT, Dutch Alpino, English EWT and Norwegian Bokmål Universal Dependencies Models (Nivre et al., 2018).

After NLP parsing, pre-processing continues with the removal of irrelevant articles, such as articles about sports and cultural events, weather forecasts, etc.[3].] The reason for removing these articles is that they often contain nonnatural language (e.g. solutions to crossword puzzles, sports results and weather forecasts), which can interfere with the construction of a word embedding model. The resulting set of processed articles is used in two ways: 1) to construct a word embedding model that in turn is used to create the sentiment dictionary, and 2) to extract sentences to validate and optimize the final sentiment dictionary.

Creating the word embedding model

In the second step, the lemmas and UPOS tags from the pre-processed articles are used to create GloVe word embedding models (Pennington et al., 2014) for each of the languages. The parameters used to generate these models are kept the same as the ones used by Rheault et al. (2016), because the goal of the study is to replicate their approach in a different domain and in different languages. Another reason for not optimizing the model parameters further is because

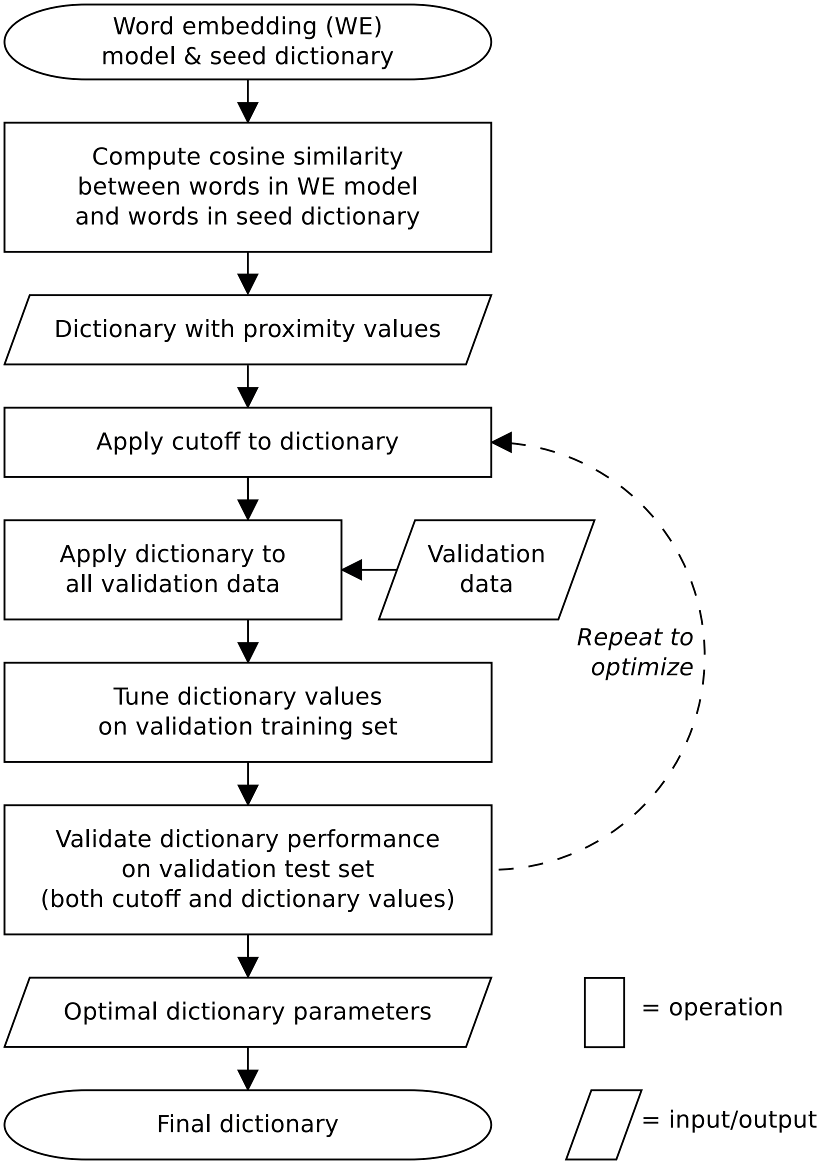
Word embedding (WE) model & seed dictionary
Compute cosine similarity between words in WE model and words in seed dictionary
Dictionary with proximity values
Apply cutoff to dictionary
Apply dictionary to all validation data
Validation data
Repeat to optimize
Tune dictionary values on validation training set
Validate dictionary performance on validation test set (both cutoff and dictionary values)
Optimal dictionary parameters
Final dictionary
= operation
= input/output

Figure 1. Data flow chart

of the computational complexity of estimating these models, especially in four languages. The parameters used are as follows: 1) tokens (a group of characters separated by whitespaces) that occur less than 5 times in the corpus are filtered out, 2) the symmetric token window size is set to 7 tokens,

Table 2. Corpora details

	Danish	Dutch[1]	English	Norwegian
Documents (x million)	1.15	2.39	2.25	1.31
Tokens (x million)	463.08	891.75	896.39	442.35
Vocabulary (x 100,000)	799.78	1176.31	615.23	763.58

Note:
[1] Both Dutch and Flemish data

meaning that the 7 tokens preceding and the 7 tokens following a token are considered as co-occurrences, 3) tokens are positioned in a 300-dimensional vector-space. Based on these model parameters, the word embedding model is estimated over 100 iterations, after which tokens occurring less than 20 times in the corpus are removed from the models. Corpus stastistics (with all tokens included) for the different languages are presented in table 2.

Creating validation data

In the third step, the articles that remain after pre-processing are filtered. Only political news articles are kept to answer the research questions and test the hypotheses, by removing any articles that do not mention at least one political actor. This is done by querying the articles at sentence level for the presence of both parties and/or individual actors (MPs, party leaders and (prime) ministers). The queries are date-limited, so that actors are only included on dates that they were actually active. Details relating to the political actor queries, and how they were constructed and executed, can be found in the appendix. All queries combined result in a total of 264,141 (BE), 309,701 (DK), 247,702 (NL), 237,244 (NO) or 512,180 (UK) newspaper articles in which one or more political actors are mentioned.

A random subset of these articles is sampled for individual sentences containing one or more political actors. These sentences are manually coded to construct a validation dataset for the sentiment dictionary.[4] They are coded by student assistants based on the following question: 'How would you describe the overall tone expressed in this sentence?' The answer options are negative, neutral/absent and positive. Training and intercoder reliability testing is done by the principal researcher in each country/language, and one (or two, in Denmark) student coders per language. Note that the coders are explicitly instructed to evaluate the valence of the sentence (its connotation), rather than the stance held towards or by the actor. During the final intercoder reliability test, 50 sentences are coded by both the researcher and the student assistant(s), after which the student codes the remaining

Table 3. Validation details

	Danish	Dutch[1]	English	Norwegian
Hand-coded sentences	3187	3538	4569	3933
Coding time[2]	7	7	14	8
Intercoder reliability[3]	0.75	0.84	0.71	0.79

Note:
[1] Both Dutch and Flemish data
[2]Median time per sentence, in seconds
[3]Using Krippendorff's alpha

sentences. English is an exception, as in contrast to the other languages sentences are not coded by native speakers, but rather Norwegian coders that are fluent in English. In table 3 the number of coded sentences, median coding time per sentence, and the intercoder reliablity are shown for each language. All student coders have been paid for their work.

Expanding the dictionary

In the fourth step, a measure is constructed to indicate the proximity of all words in the word embedding model to the seed dictionary. The seed dictionary is taken directly from Rheault et al. (2016), to replicate their method in the domain of political news. As the original seed dictionary is only in English, it is manually translated to the other three languages. The main goal during this translation process is to stay as close as possible to the original literal meaning of the English seed words as possible. While this ensures that the seed dictionaries in different languages are as similar as possible in a literal sense, it also opens up room for differences in the semantic meaning of the translations. This tradeoff is considered worthwhile, as the current process requires comparatively little human labor. In addition, recent work by Proksch et al. (2019) shows that automatic (Google Translate) translations of sentiment dictionaries perform remarkably well, illustrating the limited impact of literal translations. The full seed dictionaries for all four languages can be found in the appendix.[5]

In figure 1, the (translated) seed dictionaries and word embedding models are used as input to cacluclate the proximity of all corpus words (including the seed words themselves) to the words in the seed dictionary. Proximity is determined for each possible pair of corpus and seed words individually, by computing the cosine similarity of all pairs based on their values on the 300 dimensions of the word embedding model. By subtracting the sum of cosine similarity with the negative seed words from the sum of similarity with the positive seed words a measure is constructed indicating the relative proximity of each word to the positive and negative words in the seed dictionary. These raw values are scaled, but in a slightly different way than Rheault et al. (2016) propose. Rather than scaling the positive and negative values separately, which disregards the relative proximity of positive and negative words to the seed words, all values are scaled by dividing by the maximum absolute value among those values. This results in a dictionary of all words in the corpus with their proximity to the seed dictionary (third step in figure 1). These proximity values are operationalized as sentiment scores, as higher positive values indicate closer proximity to the positive seed words, and higher negative values indicate closer proximity to the negative seed words. Because the values are based on proximity, they also take into account context-related issues, such as negation (i.e. a positive word that is often negated will have a lower proximity to the positive seed words than a positive word that is hardly ever negated).

Creating and validating the final dictionary

In step five, the final sentiment dictionary is created by selecting words from the expanded dictionary created in the previous step, and tuning the interpretation of the proximity values. This process corresponds to all operations following the ‘Dictionary with proximity values’ input/output in figure 1. Various values, ranging from .15 to .35 in steps of .05, are tested as threshold for the minimum absolute proximity above which words are included in the dictionary. Based on the resulting dictionary, sentiment scores are computed by summing the proximity values of all sentiment words present in a sentence, and dividing that by the total number of words in the sentence. Then, these values are interpreted according to an ordinal scale (negative, neutral, positive), to make them correspond to the manual coding. The cutoff points required to convert the sentence values to ordinal categories are optimized as well, by testing cutoff values between -.1 and .1 in steps of .005 for both the positive and negative cutoff. A simplified example of how the final dictionary is constructed and applied can be found in the appendix.

As both the (absolute) proximity threshold, positive cutoff and negative cutoff are tested concurrently, a total of 8405 parameter combinations is tested using a 5-fold cross-validation approach. The hand-coded sentences are split into five equal parts/folds which are each used once to test the performance of the optimal dictionary parameters, while the other four folds are used to learn the optimal parameters. Thus, the optimal parameters are determined five times on different parts of the validation data. Similarly, the performance of those different parameter sets is also tested each time, and each time on a different part of the total hand-coded dataset.

The optimal parameters for each fold are determined based on the weighted (by the proportion of manually coded sentences in each category) F_1-score. The final performance is determined by taking the mean of all performance indicators over the 5 folds, and used to answer RQ_1 and RQ_2. Then, the optimal parameters are determined based on the whole hand-coded dataset. These parameters are used to classify sentiment for all political news articles (i.e. articles that mention at least one political actor). The sentiment scores for each sentence are aggregated to the document level, and used to test H_1 and H_2.

Summary & Costs

While the method described above is quite specific and elaborate, these steps can be generalized to a substantial extent. Most importantly, the method can be used to classify different aspects of texts than sentiment (e.g. topics), simply by using seed dictionaries with different words (see also Amsler, 2020). Assuming a corpus of texts, a word embedding model (ideally constructed from the corpus), a seed dictionary and a validation dataset, there are only two steps required to construct the final dictionary (see also figure 1). 1) Determine the optimal proximity value above which words should be included in the dictionary, and 2) determine how high the sum of the word values needs to be in order to consider a concept (e.g. topic, frame, etc.) as being present in a text/document. Both should ideally be done by using a human-labeled validation set. One might notice these steps are somewhat different from the procedure above, where positive and negative sentiment is measured using a single seed dictionary. However, two separate seed dictionaries are effectively used, and the proximity to the negative seed dictionary is subtracted from the proximity to the positive seed dictionary, to construct a single measure. This can be done with any one-dimensional concept.

To estimate the word embedding models for this study, a 16-core server with 32GB of RAM was used. The models took in total around 34 hours to estimate. The costs to compute these models was around $3.40. Of course, hand-coding of sentences used for validation is significantly more expensive. The median time for student assistants to code one sentence is between 7 and 14 seconds (see table 2). Rounding the coding time per sentence up to 15 seconds, it would take ~17 hours to code 4000 sentences per country. When including an additional 10 hours per language for training the student assistants, and assuming a wage of $15 per hour, the total costs of hand-coding

Table 4. Dictionary parameters and size

	Proximity[1]	Positive sentiment[2]	Negative sentiment[2]	Negative words	Positive words	Total words
Danish	0.30	0.03	0	11352	10211	21563
Dutch	0.30	0.04	0.005	12911	13690	26601
English	0.20	0.03	0.005	12442	10458	22900
Norwegian	0.25	0.05	0.010	14149	13994	28143

Note:
Zero and values between the positive and negative sentiment cutoffs are interpreted as neutral
[1] Minimum required proximity between word and positive/negative seed dictionary
[2] Values above/below which sentiment is interpreted as positive/negative

all four languages is $1620. As is shown below, the costs can however be even lower, as the method described here also works with a smaller hand-coded dataset. And when coding in their native language student coders are almost twice as fast as assumed here.

Validating a computer-generated sentiment dictionary

To answer both research questions, the sentiment dictionaries in each language are optimized and validated through the use of a hand-coded validation dataset. In this process two sets of parameters are tuned: 1) the threshold for including words in the dictionary (i.e. the minimal proximity score a word must have to the seed words in order to be included in the dictionary), and 2) the positive and negative cutoffs for converting the sentiment scores to categories. The cutoffs for the minimum proximity, minimum positive sentiment score, and maximum negative sentiment score (excluding 0, which is always considered as no sentiment) are presented in the first three columns of table 4. Using the optimal proximity cutoff, the final three columns of table 4 show the number of positive and negative words, as well as the total size of the dictionaries.[6]

The stability of the three dictionary parameters is tested for all languages using smaller sample sizes of 100, 500, 1000 or 2000 sentences. While these samples in some cases result in slightly different optimal dictionary parameters, the general performance of the dictionaries remains stable when using 1000 sentences or more to optimize the parameters. Smaller sample sizes tend to produce diverging dictionary parameters, and with the exception of Danish an overestimation of

Table 5. Classification performance for Polyglot and word embedding dictionaries

	F1		**Precision**		**Recall**		**# of sentences**		
	Poly.	WE	Poly.	WE	Poly.	WE	Poly.	WE	Human
Danish									
Negative	0.46	0.61	0.57	0.67	0.39	0.57	743	922	1081
Neutral	0.54	0.69	0.61	0.64	0.48	0.74	1291	1901	1659
Positive	0.33	0.40	0.23	0.45	0.59	0.36	1153	364	447
Weighted average	*0.48*	*0.62*	*0.54*	*0.62*	*0.46*	*0.63*	*3187*	*3187*	*3187*
Dutch									
Negative	0.36	0.51	0.37	0.56	0.35	0.46	816	718	863
Neutral	0.35	0.77	0.66	0.71	0.24	0.83	817	2613	2246
Positive	0.24	0.24	0.15	0.37	0.65	0.18	1905	207	429
Weighted average	*0.34*	*0.64*	*0.53*	*0.63*	*0.32*	*0.66*	*3538*	*3538*	*3538*
English									
Negative	0.61	0.66	0.70	0.72	0.53	0.61	1485	1637	1948
Neutral	0.55	0.61	0.55	0.57	0.56	0.66	1828	2094	1812
Positive	0.44	0.48	0.36	0.47	0.56	0.49	1256	838	809
Weighted average	*0.56*	*0.61*	*0.58*	*0.62*	*0.55*	*0.61*	*4569*	*4569*	*4569*
Norwegian									
Negative	0.46	0.56	0.42	0.62	0.52	0.51	1507	989	1207
Neutral	0.52	0.71	0.65	0.65	0.44	0.78	1420	2526	2108
Positive	0.37	0.39	0.30	0.49	0.49	0.33	1006	418	618
Weighted average	*0.48*	*0.61*	*0.52*	*0.62*	*0.47*	*0.63*	*3933*	*3933*	*3933*

Note: Poly. and WE refer to the Polyglot and word embedding dictionaries respectively

the dictionary performance. Upsampling is explored as a method to counteract the obvious class imbalances when classifying newspaper sentiment. However, without any exceptions, the dictionaries constructed using upsampling resulted in worse average performance (weighted F1-score) than when using dictionaries based on unbalanced samples. Thus, balancing the classes in this way does not increase the performance of the method.[7]

In table 5, the mean sentiment performance metrics for the word embedding (WE) dictionaries are shown per language, alongside the performance of the respective Polyglot dictionaries.[8] As the results show, the WE dictionaries in different languages perform comparably on average, despite the differences in size and balance between the positive and negative words. The performance of individual categories in each language is clearly related to their prevalence (i.e. the most frequently occurring category performs best, the least frequently occurring category

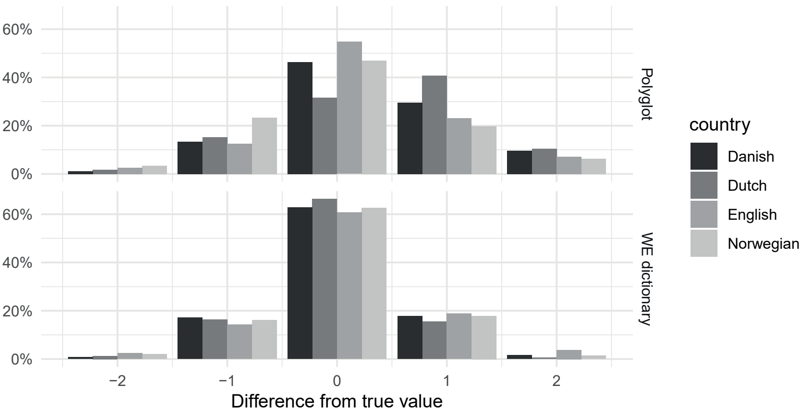

Figure 2. Difference between predicted and true sentiment category

performs worst). Generally speaking, the neutral category thus performs best, followed by the negative category, and then the positive category. The one exception in this is English, where the negative category is both slightly larger and performs slightly better than the neutral category. The weighted average F_1 scores for each of the languages, ranging between .61 and .64, are not high enough for detailed substantive research at the level of individual sentences. However, it can be argued that errors might cancel each other out when the sentiment of individual sentences is aggregated.

To test this assumption, figure 2 shows the distribution of errors between different classes for both the Polyglot and WE dictionaries[9]. The values indicate the number of steps/categories the predicted category is from the true value (e.g. +2 means the prediction is positive while the true value is negative, while -2 indicates the inverse). The 0 category thus shows the accuracy of the dictionaries. In all cases, the accuracy of the WE dictionaries is above 60%, while the Polyglot dictionaries fail to reach higher than 50% accuracy, with the exception of the English dictionary (55%). Besides making less mistakes in general, the severity of the mistakes is also lower with the WE dictionaries than with Polyglot. The vast majority of the WE errors fall within the +/-1 categories, indicating that errors between positive/negative and neutral are most common. For Polyglot, there is also a substantial amount of errors that falls in the +2 category. The distribution of errors is also less skewed for the WE dictionaries than for Polyglot, indicating that errors will cancel each other out to a larger extent in the former than in the latter. These results support the assumption that aggregation will improve the performance of the method. So while the weighted F1-scores of the WE dictionaries are too low for detailed analyses, the method is suitable for aggregate-level analyses, providing a clear answer to *RQ_1*.

Based on the average F_1 scores, the WE dictionaries outperform the Polyglot dictionaries by a substantial margin in all languages, except English. In the latter case, the performance advantage of the WE dictionary is still present, but less pronounced. Looking at the F_1 scores of individual categories, the same picture emerges, regardless of the language. Only in the Dutch positive category does the Polyglot dictionary perform on-par with the WE dictionary, and both perform equally bad. In general, the difference in performance between Polyglot and the WE dictionaries is smallest for the positive categories. This is caused primarily by the recall of the positive category being higher for Polyglot than the WE dictionaries in all languages, meaning that Polyglot captures a larger percentage of the human-coded positive sentences. This is however the only point where Polyglot outperforms the WE dictionaries. These results provide a clear answer to *RQ_2*, as the WE dictionaries perform substantially better than the alternatives provided by Polyglot. The stable performance between languages also shows that the WE approach is especially suitable for comparative research.

Investigating actor sentiment

The predictive validity of the WE sentiment dictionaries is evaluated by testing for the well-established presence of negativity bias in political news (*H_1*), and the hypothesis that this bias is stronger in tabloid newspapers than in broadsheet newspapers (*H_2*). For each country, figure 3 shows the average sentiment over time in tabloid and broadsheet newspapers (see table 1 for details). All plots are smoothed using a LOESS function with a span of .25 and the gray bands indicating the 95% confidence interval of the standard error. Descriptive statistics of the sentiment scores for the different newspapers can be found in the appendix. The sentiment shown in figure 3 is negative throughout the whole period in all countries, as is illustrated by the y-axis not reaching higher than -.1. Unsurprisingly, the mean sentiment scores per newspaper (see appendix) are also negative in all cases. Both results provide clear evidence for the presence of a negativity bias in political news, confirming *H_1*.

The graphs in figure 3 also show that the tabloid newspapers are generally speaking more negative in their coverage than the broadsheets. This difference is most pronounced in the UK, while it is moderate in Denmark,

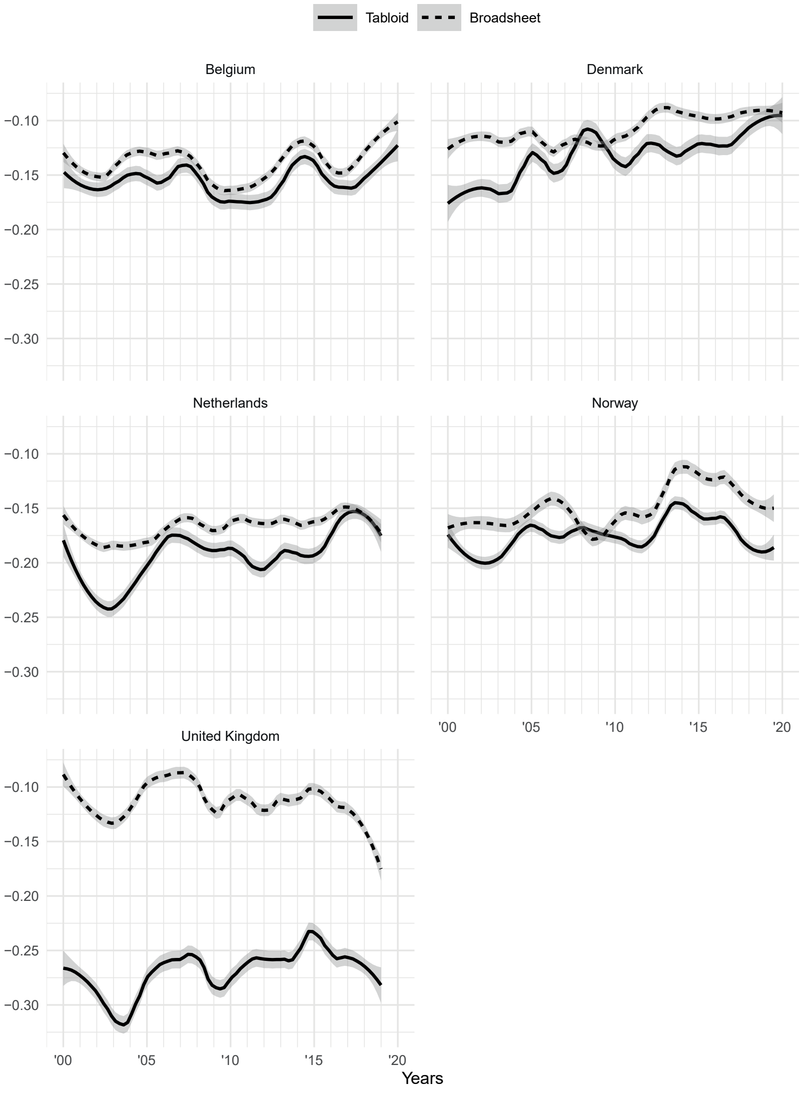

Figure 3. Sentiment by newspaper

The Netherlands and Norway. In Belgium, the difference in sentiment between the tabloid and broadsheets is still significant, but limited in size. These interpretations of the graphs are supported by the unstandardized regression results in table 6. Time is included in these models as a control variable to account for over-time developments, and while it is a significant predictor, the effect sizes are negligible in all countries. The tabloid dummy is also highly significant, and its effect sizes are at least an order of magnitude larger than those of time. Even so, the effect sizes are marginal at best, with the exception of the UK, where a stronger effect is visible. This does however not impede the testing of H_2, as it only concerns the direction (and not the strength) of the effect. Therefore the significance and negative value of the tabloid variable provides enough support to confirm H_2.

That being said, the fluctuations in sentiment that are visible in figure 3 do not really align between tabloids and broadsheets, except in Belgium and, to a lesser extent, the UK. The absence of a clear relation between tabloid and broadsheet sentiment illustrates that tabloids offer substantially different content, with substantially different sentiment, than broadsheet newspapers. The stronger relationship between tabloid and broadsheet sentiment in Belgium can be explained by the highly concentrated ownership in the Flemish newspaper market. For the UK, the explanation is not as apparent, but a tentative explanation might be that in the more professionalized and market-driven media system of the UK (see Hallin & Mancini, 2004), newspapers in general follow the sentiment of the general public, with the only difference being that tabloid newspapers are more expressive/sensational in their sentiment than broadsheets. The difference in media system also provides a tentative explanation for the relatively large difference in negativity between broadsheets and tabloids in the UK, when compared to the other countries.

Regardless of these differences, the impact of the start of the economic crisis around 2008 is clearly visible in all countries, although in Denmark the impact is most visible in tabloid coverage, while in Norway it is more pronounced in the broadsheet newspaper. Besides the onset of the economic crisis, no other trends are clearly shared between the countries. At the national level, however, the most striking trend is the increase in negativity following the Brexit referendum in the UK. The fact that both the economic crisis and Brexit referendum are clearly reflected in the sentiment of newspapers provides additional support for the validity of the sentiment measure.

Conclusion

The results presented in this paper illustrate the advantages of generating a custom sentiment dictionary based on a word embedding model and a

Table 6. Regression results by country

	Dependent variable:				
			Sentiment		
	Belgium	Denmark	Netherlands	Norway	UK
Tabloid (dummy)	−.0166***	−.0255***	−.0255***	−.0241***	−.1498***
	(.0010)	(.0012)	(.0009)	(.0011)	(.0011)
Time (in years)	.0007***	.0020***	.0020***	.0016***	−.0004***
	(.0001)	(.0001)	(.0001)	(.0001)	(.0001)
Constant	−.1432***	−.1256***	−.1833***	−.1644***	−.1084***
	(.0009)	(.0010)	(.0009)	(.0012)	(.0011)
Observations	264,141	309,701	247,702	237,244	512,180
R2	.0013	.0027	.0047	.0029	.0331
Adjusted R2	.0013	.0027	.0047	.0029	.0331
Residual Std. Error	.2249	.2816	.2179	.2703	.3643
F Statistic	172.4721***	417.3835***	588.7979***	342.2341***	8,780.0050***

Note: *p<0.1; **p<0.05; ***p<0.01

limited set of seed words. As shown in the validation section, WE dictionaries perform adequately when classifying sentiment in individual sentences, when compared to human coding. This human coding is also leveraged to improve dictionary performance by optimizing the selection of words included in the dictionary, and by tuning the interpretation of the raw sentiment scores when converting them into categories. Comparing the performance of the WE dictionaries to the well-established (see e.g. Boukes et al., 2020; van Atteveldt et al., 2021) Polyglot sentiment dictionaries, it is clear that the method described in this study provides a substantial improvement, especially in languages other than English. A likely explanation for this difference can be found in the data both methods are based on. The WE dictionaries are created specifically from the data (newspaper articles) to which they are applied, while the Polyglot dictionaries are based on the more formal language of Wikipedia articles. Another advantage of the WE method is the relatively stable performance across different languages, making it especially suitable for comparative research. This conclusion is further reinforced by the correct detection of a negativity bias in political news in all five countries, which is stronger in tabloids than in broadsheets.

That being said, the performance of the custom dictionaries when compared to human coding still leaves room for improvement. Specifically machine/deep learning methods seem to be capable of outperforming the

WE dictionaries (e.g. van Atteveldt et al., 2021, p. 128, Table 2). The use of sentences as unit of analysis and the essentially random classification errors from the WE dictionaries however make it likely that the performance of these dictionaries will be higher on the document level than on the sentence level. So even though the performance might not yet be high enough for valid sentence-level analyses, the method is performing well enough when analyzing aggregated data. There are also ample options for improvement of the method. For example, using more advanced sampling techniques to deal with the inherent class imbalances in the sentiment of political news. Or using separate cutoffs for the inclusion of positive and negative dictionary words, optimizing the seed dictionary further, and investigating which words in the dictionary most often cause errors in classification. On a more fundamental level, and assuming the availability of sufficient computing power, the method can also be further optimized by explicitly validating different sets of parameters used for generating word embedding models.

While this study presents a single, weakly supervised approach to extend a (sentiment) seed dictionary, there are many related ways to expand seed dictionaries. For example, the doctoral dissertation of Michael Amsler (2020) describes a similar but far more elaborate algorithm than the one used here. Notable differences are an iterative approach to dictionary expansion, and an extensive evaluation of the cosine similarity relationships between newly suggested words and words that are already part of the dictionary. As a result, the algorithm uses the cosine similarity between individual words, rather than the similarity to the entire pool of words in a seed dictionary. What lacks in this approach, is a point where human input can be effectively leveraged. Other studies (e.g. Alba et al., 2018; Makki et al., 2014) do make use of human input to expand their dictionaries. For example by determining the words that are most similar to a seed dictionary in a word embedding model, let humans evaluate which of those most similar words are most suitable to be included in the seed dictionary, expanding the seed dictionary and then repeat the process (Alba et al., 2018). This approach differs in its application of human labor from the current study, as it directly evaluates words, rather than providing labeled examples. This has the upside of directly (instead of indirectly) evaluating dictionary words, but also comes with the downside that the approach can become quite labor-intensive when a large vocabulary needs to be evaluated. Yet another approach is suggested by Alhothali & Hoey (2017), who combine word embeddings with pre-existing semantic resources, such as WordNet. The rationale here is that datasets containing synonyms and/or antonyms can by themselves be used for dictionary expansion, and that the semantic proximity of words in a vector space can be leveraged to improve this rule-based dictionary expansion method. An upside of this approach is that it is unsupervised (like the one described by Amsler, 2020), while the reliance on external linguistic resources limits its application to languages for which such resources are available.

Although there is room for improvement, the results presented here illustrate three main points. Firstly, it is possible to analyze sentiment at sentence (instead of document) level with reasonable accuracy, illustrating the opportunities for creating more fine-grained sentiment analysis methods in the future. Secondly, the costs of the WE dictionary approach are relatively low. The costs for constructing and optimizing a dictionary for a single country remains well below $500, with the amount of required hand-coding being limited to around 2000 sentences. In addition, the computational requirements are modest, when using corpora of sizes similar to the ones used here. And while there are substantial costs associated with the construction of the corpora used in this study, those costs do not relate exclusively to the method described here. Cleaning and NLP parsing of a corpus is a worthwhile investment for all kinds of automated text analysis methods. Finally, and perhaps most importantly, the results show there is still room for dictionary-based approaches in automated sentiment analysis, and there is no longer a need to manually create such dictionaries when working with sufficiently large data sets.

Supplementary Materials

Irrelevant article coding procedure

To classify 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. Student 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 codebook. The hand-coded articles are then used as input for a multinomial 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 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. Other performance measures can be found in table 2.

Actor query construction and execution

Data for political parties is 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, 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).

Tables

Table 1. Sentiment descriptives by newspaper

Newspaper	Mean	SD	Median	N
The Daily Telegraph	-.127	.365	-.124	165829
The Guardian	-.101	.352	-.100	202538
The Sun	-.263	.380	-.285	143813
Aftenposten	-.150	.270	-.137	104679
Dagbladet	-.159	.272	-.144	65046
VG	-.187	.269	-.179	67519
De Morgen	-.142	.217	-.131	91661
De Standaard	-.132	.220	-.120	103484
Het Laatste Nieuws	-.153	.242	-.138	68996
Ekstra Bladet	-.131	.291	-.120	67308
Jyllands-Posten	-.106	.285	-.099	133985
Politiken	-.110	.272	-.101	108408
NRC Handelsblad	-.167	.201	-.155	87718
De Telegraaf	-.188	.247	-.183	78311
De Volkskrant	-.163	.207	-.150	81673

Table 2. Irrelevant articles classification performance

	English	Norwegian	Danish	Dutch (BE)	Dutch (NL)
Accuracy	0.873	0.859	0.843	0.865	0.866
Kappa	0.737	0.715	0.685	0.730	0.731
Sensitivity	0.853	0.767	0.801	0.813	0.796
Specificity	0.906	0.944	0.883	0.917	0.934
Pos Pred Value	0.937	0.926	0.865	0.909	0.923
Neg Pred Value	0.788	0.814	0.825	0.828	0.822
Precision	0.937	0.926	0.865	0.909	0.923
Recall	0.853	0.767	0.801	0.813	0.796
F1	0.893	0.839	0.832	0.859	0.855
Prevalence	0.623	0.480	0.484	0.505	0.498
Detection Rate	0.531	0.368	0.387	0.411	0.397
Detection Prevalence	0.567	0.397	0.448	0.452	0.430
Balanced Accuracy	0.879	0.855	0.842	0.865	0.865

Table 3. Optimal dictionary parameters with various hand-coded sample sizes (Norway)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weigthed F1
100	0.20	0.030	0.005	0.6611382
500	0.25	0.035	0.010	0.6338293
1000	0.25	0.050	0.010	0.6212897
2000	0.25	0.050	0.010	0.6259237
3933	0.25	0.050	0.010	0.6141338

Table 4. Optimal dictionary parameters with various hand-coded sample sizes (UK)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weigthed F1
100	0.15	0.045	-0.010	0.6531431
500	0.20	0.035	-0.005	0.6319481
1000	0.25	0.030	0.005	0.6132555
2000	0.20	0.030	0.005	0.6172885
4569	0.20	0.030	0.005	0.6087228

Table 5. Optimal dictionary parameters with various hand-coded sample sizes (DK)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weigthed F1
100	0.30	0.015	0.00	0.6096293
500	0.30	0.030	0.00	0.6175872
1000	0.25	0.035	0.00	0.6154109
2000	0.25	0.050	0.01	0.6181187
3187	0.30	0.030	0.00	0.6222762

Table 6. Optimal dictionary parameters with various hand-coded sample sizes (NL)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weigthed F1
100	0.25	0.045	-0.005	0.7619048
500	0.25	0.060	-0.010	0.6247226
1000	0.30	0.030	0.005	0.6284429
2000	0.30	0.030	-0.005	0.6346448
3538	0.30	0.040	0.005	0.6406772

Table 7. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Norwegian)

	-1	0	1
-1	613	298	78
0	541	1648	337
1	53	162	203

Table 8. Confusion matrix (Polyglot), predictions in rows (Norwe- gian)

	-1	0	1
-1	630	740	137
0	325	917	178
1	252	451	303

Table 9. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (English)

	-1	0	1
-1	1182	346	109
0	592	1196	306
1	174	270	394

Table 10. Confusion matrix (Polyglot), predictions in rows (English)

	-1	0	1
-1	1040	330	115
0	580	1008	240
1	328	474	454

Table 11. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Danish)

	-1	0	1
-1	615	284	23
0	415	1224	262
1	51	151	162

Table 12. Confusion matrix (Polyglot), predictions in rows (Danish)

	-1	0	1
-1	420	285	38
0	356	791	144
1	305	583	265

Table 13. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Dutch)

	-1	0	1
-1	401	271	46
0	441	1866	306
1	21	109	77

Table 14. Confusion matrix (Polyglot), predictions in rows (Dutch)

	-1	0	1
-1	301	450	65
0	190	541	86
1	372	1255	278

Table 15. Sentiment classification performance (Norwegian, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.53	0.48	0.60	25	20
Neutral	0.77	0.77	0.77	57	57
Positive	0.49	0.56	0.43	18	23
Combined	0.66	0.66	0.67	100	100

Table 16. Sentiment classification performance (English, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.64	0.53	0.81	32	21
Neutral	0.72	0.83	0.63	46	60
Positive	0.54	0.50	0.58	22	19
Combined	0.65	0.66	0.68	100	100

Table 17. Sentiment classification performance (Danish, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.58	0.51	0.68	41	31
Neutral	0.67	0.72	0.63	46	52
Positive	0.47	0.54	0.41	13	17
Combined	0.61	0.61	0.62	100	100

Table 18. Sentiment classification performance (Dutch, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.57	0.50	0.67	8	6
Neutral	0.86	0.87	0.84	31	32
Positive	0.29	0.33	0.25	3	4
Combined	0.76	0.76	0.77	42	42

Table 19. Sentiment classification performance (Norwegian, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.56	0.51	0.62	141	117
Neutral	0.74	0.78	0.69	274	311
Positive	0.43	0.40	0.47	85	72
Combined	0.63	0.64	0.63	500	500

Table 20. Sentiment classification performance (English, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.65	0.57	0.76	194	146
Neutral	0.66	0.73	0.60	212	258
Positive	0.53	0.53	0.52	94	96
Combined	0.63	0.63	0.65	500	500

Table 21. Sentiment classification performance (Danish, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.60	0.54	0.68	181	145
Neutral	0.68	0.73	0.64	251	285
Positive	0.42	0.43	0.41	68	70
Combined	0.62	0.62	0.62	500	500

Table 22. Sentiment classification performance (Dutch, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.54	0.49	0.60	145	119
Neutral	0.74	0.82	0.68	294	357
Positive	0.26	0.18	0.46	61	24
Combined	0.62	0.65	0.63	500	500

Table 23. Sentiment classification performance (Norwegian, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.55	0.50	0.61	287	236
Neutral	0.73	0.81	0.67	546	664
Positive	0.37	0.30	0.50	167	100
Combined	0.62	0.64	0.62	1000	1000

Table 24. Sentiment classification performance (English, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.63	0.57	0.72	418	331
Neutral	0.63	0.68	0.59	411	476
Positive	0.52	0.55	0.49	171	193
Combined	0.61	0.61	0.63	1000	1000

Table 25. Sentiment classification performance (Danish, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.57	0.50	0.67	337	253
Neutral	0.70	0.75	0.65	534	620
Positive	0.38	0.38	0.39	129	127
Combined	0.62	0.62	0.62	1000	1000

Table 26. Sentiment classification performance (Dutch, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.53	0.51	0.57	281	250
Neutral	0.73	0.77	0.69	590	663
Positive	0.37	0.31	0.46	129	87
Combined	0.63	0.64	0.63	1000	1000

Table 27. Sentiment classification performance (Norwegian, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.56	0.51	0.62	594	492
Neutral	0.73	0.80	0.67	1109	1313
Positive	0.37	0.30	0.46	297	195
Combined	0.63	0.64	0.63	2000	2000

Table 28. Sentiment classification performance (English, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.65	0.59	0.71	817	677
Neutral	0.63	0.69	0.58	822	970
Positive	0.51	0.51	0.52	361	353
Combined	0.62	0.62	0.63	2000	2000

Table 29. Sentiment classification performance (Danish, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.61	0.61	0.62	660	649
Neutral	0.69	0.73	0.65	1061	1192
Positive	0.35	0.28	0.48	279	159
Combined	0.62	0.63	0.62	2000	2000

Table 30. Sentiment classification performance (Dutch, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.52	0.48	0.56	520	445
Neutral	0.75	0.80	0.70	1233	1412
Positive	0.32	0.26	0.44	247	143
Combined	0.63	0.65	0.63	2000	2000

Table 31. Sentiment classification performance (Norwegian, with upsampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.55	0.52	0.60	1207	1048
Neutral	0.65	0.61	0.69	2108	1868
Positive	0.42	0.55	0.34	618	1017
Combined	0.58	0.57	0.61	3933	3933

Table 32. Sentiment classification performance (English, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.66	0.61	0.72	1948	1657
Neutral	0.59	0.58	0.60	1812	1737
Positive	0.49	0.60	0.41	809	1175
Combined	0.60	0.60	0.62	4569	4569

Table 33. Sentiment classification performance (Danish, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.62	0.61	0.64	1081	1027
Neutral	0.62	0.57	0.68	1659	1396
Positive	0.43	0.58	0.34	447	764
Combined	0.59	0.58	0.62	3187	3187

Table 34. Sentiment classification performance (Dutch, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.48	0.43	0.55	863	670
Neutral	0.69	0.65	0.73	2246	1999
Positive	0.34	0.51	0.25	429	869
Combined	0.60	0.58	0.63	3538	3538

Table 35. Norwegian seed dictionary

Positive seed words		**Negative seed words**	
dyktig_ADJ	glede_NOUN	misbruk_NOUN	fryktelig_ADJ
beundringsverdig_ADJ	vennligst_ADJ	redd_ADJ	såre_VERB
verdsette_VERB	elske_VERB	sinne_NOUN	uvel_ADJ
hensiktsmessig_ADJ	herlig_ADJ	sint_ADJ	mangelfull_ADJ
vakker_ADJ	kjærlig_ADJ	angst_NOUN	utilstrekkelig_ADJ
beste_ADJ	glimrende_ADJ	bekymre_ADJ	mindreverdig_ADJ
bedre_VERB	fordel_NOUN	dårlig_ADJ	urettferdighet_NOUN
klok_ADJ	snill_ADJ	brudd_NOUN	irrelevant_ADJ
støtte_NOUN	perfekt_ADJ	brutal_ADJ	miste_VERB
komfortabel_ADJ	perfeksjon_NOUN	byrde_NOUN	tap_NOUN
sikker_ADJ	behagelig_ADJ	uforsiktig_ADJ	elendig_ADJ
kreativ_ADJ	ros_NOUN	klage_VERB	tabbe_NOUN
fryd_NOUN	skikkelig_ADJ	klage_NOUN	forsømme_VERB
hyggelig_ADJ	velstand_NOUN	forvirring_NOUN	tull_NOUN
ønskelig_ADJ	beskytte_VERB	forakt_NOUN	smerte_NOUN
verdighet_NOUN	fornuftig_ADJ	korrupt_ADJ	smertefull_ADJ
virkningsfull_ADJ	pålitelig_ADJ	korrupsjon_NOUN	dårlig_PROPN
effektivitet_NOUN	respekt_NOUN	kritikk_NOUN	fordom_NOUN
effektiv_ADJ	respektere_VERB	skade_NOUN	problem_NOUN
oppmuntre_VERB	trygg_ADJ	fare_NOUN	beklagelse_NOUN
nyte_VERB	tilfredshet_NOUN	farlig_ADJ	innskrenke_VERB
utmerket_ADJ	tilfredsstille_ADJ	død_NOUN	restriksjon_NOUN
rettferdig_ADJ	tilfredsstille_VERB	ødelegge_VERB	latterlig_ADJ
åpen_ADJ	sikre_VERB	vanskelig_ADJ	risiko_NOUN
gunstig_ADJ	betydningsfull_ADJ	vanskelighet_NOUN	trist_ADJ
heldigvis_ADV	oppriktig_ADJ	ulempe_NOUN	skam_NOUN
frihet_NOUN	smart_ADJ	skuffelse_NOUN	syk_ADJ
vennlig_ADJ	løsning_NOUN	ulykke_NOUN	dum_ADJ
vennskap_NOUN	flott_ADJ	katastrofal_ADJ	lide_VERB
oppnå_VERB	styrke_NOUN	ubehag_NOUN	forferdelig_ADJ
sjenerøs_ADJ	forsterke_VERB	nød_NOUN	trussel_NOUN
ekte_ADJ	sterk_ADJ	fiende_NOUN	tragedie_NOUN

fornøyd_ADJ	lykkes_VERB	feil_NOUN	tragisk_ADJ
vidunderlig_ADJ	suksess_NOUN	ond_ADJ	stygg_ADJ
god_ADJ	vellykket_ADJ	overdrivelse_NOUN	uønsket_ADJ
takknemlig_ADJ	suveren_ADJ	overdreven_ADJ	urimelig_ADJ
lykke_NOUN	sympatisk_ADJ	mislykkes_VERB	uheldig_ADJ
glad_ADJ	sympati_NOUN	fiasko_NOUN	dessverre_ADV
sunn_ADJ	talent_NOUN	falsk_ADJ	mislykket_ADJ
hjelpe_VERB	sann_ADJ	mangel_NOUN	urettferdig_ADJ
hjelpsom_ADJ	genuint_ADJ	frykte_NOUN	irrasjonell_ADJ
ærlig_ADJ	sannhet_NOUN	engstelig_ADJ	uakseptabel_ADJ
ære_NOUN	nyttig_ADJ	svindel_NOUN	svak_ADJ
viktighet_NOUN	verdifull_A1D2 J	skremme_VERB	svakhet_NOUN
viktig_ADJ	sprek_ADJ	ubehagelig_ADJ	hensynsløs_ADJ
forbedre_VERB	velkommen_ADJ	skade_VERB	bekymre_VERB
bedring_NOUN	bra_ADJ	skadelig_ADJ	dårligere_ADJ
integritet_NOUN	lur_ADJ	hate_VERB	dårligst_ADJ
intelligent_ADJ	fantastisk_ADJ	hat_NOUN	ynkelig_ADJ
interessant_ADJ	verdig_ADJ	håpløs_ADJ	galt_ADJ

Table 36. English seed dictionary

Positive seed words		**Negative seed words**	
able_ADJ	joy_NOUN	abuse_NOUN	horrible_ADJ
admirable_ADJ	kindly_ADV	afraid_ADJ	hurt_VERB
appreciate_VERB	love_VERB	anger_NOUN	ill_ADJ
appropriate_ADJ	lovely_ADJ	angry_ADJ	imperfect_ADJ
beautiful_ADJ	loving_ADJ	anxiety_NOUN	inadequate_ADJ
best_ADJ	magnificent_ADJ	anxious_ADJ	inferior_ADJ
better_ADJ	merit_NOUN	bad_ADJ	injustice_NOUN
clever_NOUN	nice_ADJ	breach_NOUN	irrelevant_ADJ
comfort_NOUN	perfect_ADJ	brutal_ADJ	lose_VERB
comfortable_ADJ	perfection_NOUN	burden_NOUN	loss_NOUN
confident_ADJ	pleasant_ADJ	careless_ADJ	miserable_ADJ
creative_ADJ	praise_NOUN	complain_VERB	mistake_NOUN
delight_NOUN	properly_ADV	complaint_NOUN	neglect_VERB
delightful_ADJ	prosperity_NOUN	confusion_NOUN	nonsense_NOUN
desirable_ADJ	protect_VERB	contempt_NOUN	pain_NOUN
dignity_NOUN	reasonable_ADJ	corrupt_ADJ	painful_ADJ
effective_ADJ	reliable_ADJ	corruption_NOUN	poorly_ADV
efficiency_NOUN	respect_NOUN	criticism_NOUN	prejudice_NOUN
efficient_ADJ	respected_ADJ	damage_NOUN	problem_NOUN
encourage_VERB	safe_ADJ	danger_NOUN	regret_NOUN
enjoy_VERB	satisfaction_NOUN	dangerous_ADJ	restrict_VERB
excellent_ADJ	satisfactory_ADJ	death_NOUN	restriction_NOUN
fair_ADJ	satisfying_ADJ	destroy_VERB	ridiculous_ADJ

Positive seed words		Negative seed words	
fairly_ADV	secure_VERB	difficult_ADJ	risk_NOUN
fortunate_ADJ	significant_ADJ	difficulty_NOUN	sad_ADJ
fortunately_ADV	sincere_NOUN	disadvantage_NOUN	shame_NOUN
freedom_NOUN	smart_ADJ	disappointment_NOUN	sick_ADJ
friendly_ADJ	solution_NOUN	disaster_NOUN	stupid_ADJ
friendship_NOUN	splendid_ADJ	disastrous_ADJ	suffer_VERB
gain_VERB	strength_NOUN	discomfort_NOUN	terrible_ADJ
generous_ADJ	strengthen_VERB	distress_NOUN	threat_NOUN
genuine_ADJ	strong_ADJ	enemy_NOUN	tragedy_NOUN
glad_ADJ	succeed_VERB	error_NOUN	tragic_ADJ
glorious_ADJ	success_NOUN	evil_ADJ	ugly_ADJ
good_ADJ	successful_ADJ	excess_NOUN	undesirable_ADJ
grateful_ADJ	superior_ADJ	excessive_ADJ	unfair_ADJ
happiness_NOUN	sympathetic_ADJ	fail_VERB	unfortunate_ADJ
happy_ADJ	sympathy_NOUN	failure_NOUN	unfortunately_ADV
healthy_ADJ	talent_NOUN	false_ADJ	unhappy_ADJ
help_VERB	true_ADJ	fault_NOUN	unjust_ADJ
helpful_ADJ	truly_ADV	fear_NOUN	unreasonable_ADJ
honest_ADJ	truth_NOUN	fearful_ADJ	unsatisfactory_ADJ
honour_NOUN	useful_ADJ	fraud_NOUN	weak_ADJ
importance_NOUN	valuable_AD1J3	frightened_ADJ	weakness_NOUN
important_ADJ	vigorous_ADJ	grim_ADJ	wicked_ADJ
improve_VERB	welcome_ADJ	harm_VERB	worry_VERB
improvement_NOUN	well_ADV	harmful_ADJ	worse_ADJ
integrity_NOUN	wise_ADJ	hate_VERB	worst_ADJ
intelligent_ADJ	wonderful_ADJ	hatred_NOUN	wretched_ADJ
interesting_ADJ	worthy_ADJ	hopeless_ADJ	wrong_ADV

Table 37. Danish seed dictionary

Positive seed words		Negative seed words	
dygtig_ADJ	glæde_NOUN	misbrug_NOUN	frygtelig_ADJ
beundringsværdig_ADJ	venligt_ADV	bange_ADJ	såre_VERB
værdsætte_VERB	elske_VERB	vrede_ADJ	usund_ADJ
passende_ADJ	dejlig_ADJ	vred_ADJ	mangelfuld_ADJ
smuk_ADJ	kærlig_ADJ	bekymring_NOUN	utilstrækkelig_ADJ
bedst_ADJ	storslået_ADJ	ængstelig_ADJ	lavere_ADJ
bedre_ADJ	fortjeneste_NOUN	dårlig_ADJ	uretfærdighed_NOUN
klog_ADJ	rar_ADJ	brud_NOUN	uvedkommende_VERB
trøst_NOUN	perfekt_ADJ	brutal_ADJ	tabe_VERB
komfortabel_ADJ	perfektion_NOUN	belastning_NOUN	tab_NOUN
fortrøstningsfuld_ADJ	behagelig_ADJ	uforsigtig_ADJ	elendig_ADJ
kreativ_ADJ	ros_NOUN	klage_VERB	fejl_NOUN

Positive seed words		**Negative seed words**	
fornøjelse_NOUN	ordentlig_ADV	klage_NOUN	forsømme_VERB
fornøjelig_ADJ	fremgang_NOUN	forvirring_NOUN	vrøvl_NOUN
attraktiv_ADJ	beskytte_VERB	foragt_NOUN	smerte_NOUN
værdighed_NOUN	fornuftig_ADJ	korrupt_ADJ	smertelig_ADJ
effektfuld_ADJ	pålidelig_ADJ	korruption_NOUN	elendigt_ADV
effektivitet_NOUN	respekt_NOUN	kritik_NOUN	fordom_NOUN
effektiv_ADJ	anerkendt_ADJ	ødelæggelse_NOUN	problem_NOUN
opmuntre_VERB	tryg_ADJ	fare_NOUN	beklagelse_NOUN
nyde_VERB	tilfredshed_NOUN	farlig_ADJ	begrænse_VERB
fremragende_ADJ	overbevisende_VERB	død_NOUN	restriktion_NOUN
rimelig_ADJ	tilfredsstillende_ADJ	ødelægge_VERB	latterlig_ADJ
ganske_ADV	sikre_VERB	svær_ADJ	risiko_NOUN
heldig_ADJ	betydningsfuld_ADJ	besvær_NOUN	trist_ADJ
heldigvis_ADV	oprigtig_ADJ	ulempe_NOUN	skam_NOUN
frihed_NOUN	smart_ADJ	skuffelse_NOUN	syg_ADJ
venlig_ADJ	løsning_NOUN	katastrofe_NOUN	dum_ADJ
venskab_NOUN	flot_ADJ	katastrofal_ADJ	lide_VERB
opnå_VERB	styrke_NOUN	ubehag_NOUN	forfærdelig_ADJ
gavmild_ADJ	forstærke_VERB	sorg_NOUN	trussel_NOUN
ægte_ADJ	stærk_ADJ	fjende_NOUN	tragedie_NOUN
glad_ADJ	lykkes_VERB	fejltagelse_NOUN	tragisk_ADJ
pragtfuld_ADJ	succes_NOUN	ond_ADJ	grim_ADJ
god_ADJ	vellykket_ADJ	overskridelse_NOUN	uønsket_ADJ
taknemmelig_ADJ	overlegenhed_NOUN	overdreven_ADJ	unfair_ADJ
lykke_NOUN	sympatisk_ADJ	mislykkes_VERB	ulykkelig_ADJ
lykkelig_ADJ	sympati_NOUN	nederlag_NOUN	uheldigvis_ADV
sund_ADJ	talent_NOUN	falsk_ADJ	utilfreds_ADJ
hjælpe_VERB	sand_ADJ	mangel_NOUN	uretfærdig_ADJ
hjælpsom_ADJ	virkelig_ADV	frygt_NOUN	urimelig_ADJ
ærlig_ADJ	sandhed_NOUN	frygtsom_ADJ	utilfredsstillende_ADJ
ære_NOUN	nyttig_ADJ	bedrageri_NOUN	svag_ADJ
betydning_NOUN	værdifuld_14ADJ	skræmt_ADJ	svaghed_NOUN
vigtig_ADJ	energisk_ADJ	barsk_ADJ	rædselsfuld_ADJ
forbedre_VERB	velkommen_ADJ	skade_VERB	bekymre_VERB
forbedring_NOUN	godt_ADV	skadelig_ADJ	værre_ADJ
integritet_NOUN	forstandig_ADJ	hade_VERB	værst_ADV
intelligent_ADJ	vidunderlig_ADJ	had_NOUN	stakkels_ADJ
interessant_ADJ	værdig_ADJ	håbløs_ADJ	forkert_ADJ

Table 38. Dutch seed dictionary

Positive seed words		Negative seed words	
capabel_ADJ	vreugde_NOUN	misbruik_NOUN	verschrikkelijk_ADJ
bewonderenswaardig_ADJ	welwillend_ADJ	bevreesd_ADJ	kwetsen_VERB
waarderen_VERB	liefhebben_VERB	woede_NOUN	kwalijk_ADJ
passend_ADJ	lief_ADJ	woedend_ADJ	imperfect_ADJ
mooi_ADJ	liefdevol_ADJ	ongerustheid_NOUN	ontoereikend_ADJ
best_ADJ	prachtig_ADJ	bezorgd_ADJ	inferieur_ADJ
beter_ADJ	verdienste_NOUN	slecht_ADJ	onrecht_NOUN
slim_NOUN	prettig_ADJ	breuk_NOUN	onbelangrijk_ADJ
comfort_NOUN	perfect_ADJ	wreed_ADJ	verliezen_VERB
comfortabel_ADJ	perfectie_NOUN	last_NOUN	verlies_NOUN
overtuigd_ADJ	aangenaam_ADJ	onzorgvuldig_ADJ	miserabel_ADJ
creatief_ADJ	lof_NOUN	klagen_VERB	vergissing_NOUN
genot_NOUN	juist_ADV	klacht_NOUN	verwaarlozen_VERB
verrukkelijk_ADJ	voorspoed_NOUN	verwarring_NOUN	nonsens_NOUN
wenselijk_ADJ	beschermen_VERB	minachting_NOUN	pijn_NOUN
waardigheid_NOUN	redelijk_ADJ	corrupt_ADJ	pijnlijk_ADJ
effectief_ADJ	betrouwbaar_ADJ	corruptie_NOUN	slecht_ADV
efficiëntie_NOUN	respect_NOUN	kritiek_NOUN	vooroordeel_NOUN
efficiënt_ADJ	geliefd_ADJ	schade_NOUN	probleem_NOUN
aanmoedigen_VERB	veilig_ADJ	gevaar_NOUN	spijt_NOUN
genieten_VERB	voldoening_NOUN	gevaarlijk_ADJ	beperken_VERB
uitstekend_ADJ	voldoende_ADJ	dood_NOUN	beperking_NOUN
eerlijk_ADJ	bevredigend_ADJ	vernietigen_VERB	belachelijk_ADJ
tamelijk_ADV	beveiligen_VERB	moeilijk_ADJ	risico_NOUN
fortuinlijk_ADJ	significant_ADJ	moeilijkheid_NOUN	verdrietig_ADJ
gelukkig_ADJ	oprecht_ADJ	nadeel_NOUN	schaamte_NOUN
vrijheid_NOUN	slim_ADJ	teleurstelling_NOUN	ziek_ADJ
vriendelijk_ADJ	oplossing_NOUN	ramp_NOUN	dom_ADJ
vriendschap_NOUN	schitterend_ADJ	rampzalig_ADJ	lijden_VERB
winnen_VERB	kracht_NOUN	ongemak_NOUN	vreselijk_ADJ
vrijgevig_ADJ	versterken_VERB	nood_NOUN	bedreiging_NOUN
authentiek_ADJ	sterk_ADJ	vijand_NOUN	tragedie_NOUN
verheugd_ADJ	slagen_VERB	fout_NOUN	tragisch_ADJ
glorieus_ADJ	succes_NOUN	onheil_ADJ	lelijk_ADJ
goed_ADJ	succesvol_ADJ	overdaad_NOUN	onwenselijk_ADJ
dankbaar_ADJ	superieur_ADJ	overdadig_ADJ	oneerlijk_ADJ
geluk_NOUN	sympathiek_ADJ	falen_VERB	onfortuinlijk_ADJ
blij_ADJ	sympathie_NOUN	mislukking_NOUN	helaas_ADV
gezond_ADJ	talent_NOUN	onjuist_ADJ	ongelukkig_ADJ
helpen_VERB	waar_ADJ	schuld_NOUN	onrechtvaardig_ADJ
behulpzaam_ADJ	werkelijk_ADJ	angst_NOUN	onredelijk_ADJ
oprecht_ADJ	waarheid_NOUN	angstig_ADJ	onbevredigend_ADJ

Positive seed words		Negative seed words	
eer_NOUN	bruikbaar_ADJ	fraude_NOUN	zwak_ADJ
belang_NOUN	waardev1o5l_ADJ	bang_ADJ	zwakte_NOUN
belangrijk_ADJ	krachtig_ADJ	grimmig_ADJ	goddeloos_ADJ
verbeteren_VERB	welkom_ADJ	schaden_VERB	piekeren_VERB
verbetering_NOUN	goed_ADV	schadelijk_ADJ	slechter_ADJ
integriteit_NOUN	wijs_ADJ	haten_VERB	slechtst_ADJ
intelligent_ADJ	geweldig_ADJ	haat_NOUN	ellendig_ADJ
interessant_ADJ	waardig_ADJ	hopeloos_ADJ	fout_ADJ

Notes

1. An annotated reproducible example of the adapted method is provided at https://github.com/vriezer/sentiment.
2. Even when avoiding ambiguous seed words, the final dictionary might still be biased due to the presence of bias in the (source data of the) WE model used to construct the dictionary.
3. A full description of the irrelevant article coding procedure and its results can be found in the appendix at https://osf.io/tb3kr/
4. The focus on political actors is due to reasons of data availability, and coders are instructed to ignore their presence when coding sentiment.
5. Dutch and Flemish are treated as a single language (Dutch), but as geographically distinct media markets/domains.
6. Replication materials to reproduce the results presented in this section are available as supplementary material at https://osf.io/tb3kr/.
7. Full performance results for all languages for both the sample size and upsampling experiments can be found in the appendix.
8. Spearman rank order correlation with human coding. WE: 0.464 (Danish), 0.347 (Dutch), 0.460 (English), 0.399 (Norwegian). Polyglot: 0.255 (Danish), 0.157 (Dutch), 0.385 (English), 0.224 (Norwegian).
9. Confusion matrices can be found in the appendix.

References

Alba, A., Gruhl, D., Ristoski, P., & Welch, S. (2018). Interactive dictionary expansion using neural language models. *HumL@ ISWC*, 7–15.

Aldayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, *58* (4), 102597. https://doi.org/10.1016/j.ipm.2021.102597

Alhothali, A., & Hoey, J. (2017). Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations. *arXiv:1703.09825* [*Cs*]. https://arxiv.org/abs/1703.09825

Almeida, F., & Xexéo, G. (2019). Word Embeddings: A Survey. *arXiv:1901.09069* [*Cs, Stat*]. https://arxiv.org/abs/1901.09069

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). *Polyglot: Distributed Word Representations for Multilingual NLP*. 10.

Amsler, M. (2020). *Using Lexical-Semantic Concepts for Fine-Grained Classification in the Embedding Space* [PhD thesis]. University of Zurich.

Bleich, E., & van der Veen, A. M. (2018). Media portrayals of Muslims: A comparative sentiment analysis of American newspapers, 1996–2015. *Politics, Groups, and Identities*, 1–20. https://doi.org/10.1080/21565503.2018.1531770

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14* (2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Chen, Y., & Skiena, S. (2014). Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), 383–389. https://doi.org/10.3115/v1/P14-2063

de Vreese, C., Esser, F., & Hopmann, D. N. (2016). *Comparing Political Journalism*. Routledge. https://doi.org/10.4324/9781315622286

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Glogger, I. (2019). Soft Spot for Soft News? Influences of Journalistic Role Conceptions on Hard and Soft News Coverage. *Journalism Studies*, *20* (16), 2293–2311. https://doi.org/10.1080/1461670X.2019.1588149

Hallin, D. C., & Mancini, P. (2004). Comparing Media Systems: Three Models of Media and Politics. In *Cambridge Core*. /core/books/comparing-mediasystems/B7A12371782B7A1D62BA1A72C1395E43; Cambridge University Press. https://doi.org/10.1017/CBO9780511790867

Hlavac, M. (2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* https://CRAN.R-project.org/package=stargazer.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44* (4), 491–511. https://doi.org/10.1177/0165551517703514

Lengauer, G., Esser, F., & Berganza, R. (2012). Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism*, *13* (2), 179–202. https://doi.org/10.1177/1464884911427800

Makki, R., Brooks, S., & Milios, E. E. (2014). Context-specific sentiment lexicon expansion via minimal user interaction. *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, 178–186.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. Meiselman (Ed.), *Emotion measurement*. Elsevier.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries. *Political Communication*, *36* (2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., & Zhu, H. (2018). *Universal dependencies 2.3*.

Otto, L., Glogger, I., & Boukes, M. (2017). The Softening of Journalistic Political Communication: A Comprehensive Framework Model of Sensationalism, Soft News, Infotainment, and Tabloidization. *Communication Theory*, *27* (2), 136–155. https://doi.org/10.1111/comt.12102

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, *44* (1), 97–131. https://doi.org/10.1111/lsq.12218

Reinemann, C., Stanyer, J., Scherr, S., & Legnante, G. (2012). Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, *13* (2), 221–239. https://doi.org/10.1177/1464884911427803

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS ONE*, *11* (12), e0168843. https://doi.org/10.1371/journal.pone.0168843

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12* (2-3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Shi, T., Malioutov, I., & İrsoy, O. (2020). Semantic Role Labeling as Syntactic Dependency Parsing. *arXiv:2010.11170* [*Cs*]. https://arxiv.org/abs/2010.11170

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, *659* (1), 108–121. https://doi.org/10.1177/0002716215569217

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*, 88–99.

van Atteveldt, W., Sheafer, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War. *Political Analysis*, *25* (02), 207–222. https://doi.org/10.1017/pan.2016.12

van Atteveldt, W., van der Velden, M. A. C. G., & 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), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29* (2), 205–231. https://doi.org/10.1080/10584609.2012.671234