Evaluating Word Expansion for Multilingual Sentiment Analysis of Parliamentary Speech

Yana Nikolaeva Nikolova, Costanza Navarretta
University of Copenhagen, Department of Nordic Studies and Linguistics
Emil Holms Kanal 2, 2300 Copenhagen, Denmark
yana.nik98@gmail.com, costanza@hum.ku.dk

Abstract

This paper replicates and evaluates the word expansion (WE) method for sentiment lexicon generation from Rheault et al. (2016), applying it to two novel corpora of parliamentary speech from Denmark and Bulgaria. GloVe embeddings and vector similarity are leveraged to expand synonym seed lists with domain-specific terms from the speech corpora. The resulting Danish and Bulgarian lexica are compared to other multilingual lexica by analyzing a gold standard of speech excerpts annotated for sentiment. WE correlates best with hand-coded annotations for Danish, while a machine-translated Lexicoder dictionary does best for Bulgarian. WE performance is also found to be very sensitive to processing and scoring techniques, though this is also an issue with the other lexica. Overall, automatic lexicon translation best balances computational complexity and accuracy across both languages, but robust language-agnosticism remains elusive. Theoretical and practical problems of WE are discussed.

Keywords: sentiment analysis, parliament, multilingual, word embeddings, word expansion, translation

1. Introduction

Automated textual analysis of parliamentary speeches has been a live field of interest in recent years (Abercrombie and Batista-Navarro, 2020) owing to the copiousness, availability and political import of this data. A major strand of this work has focused on the emotional and affective character of parliamentary speech, integrating insights and methods from the field of sentiment analysis. Sentiment information can be used to complement topic analyses (Bhatia and P, 2018, Abercrombie, 2021) or index political roles such as government and opposition (Proksch et al., 2019).

Recent efforts to create and format data sets of parliamentary speech material from different countries, like the CLARIN ParlaMint project (Erjavec et al., 2023), have facilitated cross-lingual comparisons and highlighted the need for methods of sentiment analysis that are computationally efficient and language-agnostic.

One such approach leverages word embedding similarity to create bespoke sentiment lexica from entire corpora of parliamentary speeches (Rheault et al., 2016; Hargrave and Blumenau, 2022). In this paper, we implement and evaluate this methodology, applying it to novel parliamentary data in Danish and Bulgarian in comparison with other multilingual lexicon approaches.

2. Related Work

Lexicon methods remain common in parliamentary speech analysis (Abercrombie and Batista-Navarro, 2020), despite a recent turn to neural and transformer-based approaches (Miok et al., 2022, Han, 2022), due to the computational efficiency, scalability and adaptability in their implementation. Popular general-use lexica for English sentiment analysis include ANEW (Bradley and Lang, 1999) and the Lexicoder Sentiment Dictionary (LSD) (Young and Soroka, 2012), which were both adapted from older resources in psychology. Automatic or semi-automatic strategies for compiling or translating lexicons are also common (Kapukaranov and Nakov, 2015, Chen and Skiena, 2014, Proksch et al., 2019) and can be especially useful for languages without available lexical resources.

General-use lexica and sentiment classifiers can, however, face challenges when applied to different styles, registers or genres (e.g., Aue and Gamon, 2005). Parliamentary proceedings contain terms like “health”, “security” and “opposition”, which have different sentiment connotations in a policy context compared to casual speech. This is why Rheault et al. (2016) turn to the word expansion (WE) methodology pioneered by Turney and Littman (2003) to create a sentiment lexicon which reflects domain-specific usage in order to analyze long-term changes in sentiment in the British Parliament. By creating embeddings from the Hansard corpus, Rheault et al. (2016) expand a seed list of manually compiled sentiment words to include semantically similar words from the target domain. However, despite the focus on domain-specificity, Rheault et al. (2016) only evaluate their lexicon on a set of tagged movie reviews.

More recently, Rice and Zorn (2021) have applied a similar method to English movie reviews and

LREC-COLING 2024, pages 6557–6563
20-25 May, 2024.

court judge opinions with success. Hargrave and Blumenau (2022) also analyze the British Parliament, using WE to measure not only sentiment, but also constructs such as "aggression" and "human narrative", in a study of gendered speech patterns. Vries (2022) adapts WE for newspaper articles in four Germanic languages (including English and Danish), introducing intermediate tuning and validation steps. His approach outperforms Polyglot, another automated lexicon generation method developed by Chen and Skiena (2014). For Polyglot, Chen and Skiena (2014) build a multilingual semantic network using common resources such as WordNets and Wikipedias and propagate an English sentiment list through the graph, finding all related words in 81 languages and using them as sentiment dictionaries.

However, the computational cost of approaches like graph propagation and lexical expansion is not negligible. Evaluation practices between studies also vary widely and interact with prior research in a limited manner, making it hard to compare their results. For example, Proksch et al. (2019) recently proposed machine translation as a much simpler way of constructing high-quality multilingual parliamentary sentiment lexica. By expanding the partially stemmed Lexicoder Sentiment Dictionary (Young and Soroka, 2012) and translating it, they obtain lexica in 20 languages, achieving high correlations to hand-coded sentiment.

This study aims to bring these results together, comparing three major approaches to lexicon creation (WE, graph propagation and translation) in applying them to novel datasets and languages to yield knowledge and recommendations for future practice.

3. Data

3.1. ParlaMint

ParlaMint is a project associated with the European research infrastructure CLARIN that has engaged researchers to collect and annotate parliamentary speech data according to a common standard, the Parla-CLARIN TEI-based encoding (Erjavec et al., 2023). Each ParlaMint corpus comes in two TEI-formats: plain-text and annotated. The annotated versions encode pre-processed and tokenized text with detailed linguistic coding including lemma and PoS-tagging.

The Bulgarian and Danish ParlaMint corpora (versions 3.0) were chosen for this study due to the linguistic knowledge necessary to implement WE, and because they are different enough to represent some of the range of the corpus.

ParlaMint-DK contains 40.8M tokens, while ParlaMint-BG is more modest at 26.4M tokens. The entire annotated corpora, covering proceedings from the period 2014-2022, were used to create word embedding vocabularies, so no sampling was involved.

3.2. Sentiment Annotation

In order to evaluate the performance of the sentiment analysis techniques, a set of 300 speeches excerpts from each ParlaMint corpus was randomly sampled for manual annotation by the author. Speeches were excerpted up to the segment containing the 100th word. Since segments are annotated slightly differently in the two corpora, Danish excerpts tend to be shorter than Bulgarian ones. Samples were representative with respect to party status (coalition, opposition, other), which is known to affect sentiment in parliament (Rudkowsky et al., 2018).

One third of the speech samples were also provided to a secondary annotator to assess reliability (Table 1). Speeches were annotated on a 5-point scale which distinguishes two degrees of emotional polarity in each direction. Inter-annotator reliability was assessed for the full scale and a simplified 3-point scale (positive, neutral, negative).

Parl.	κ_3	κ_5
BG	0,78 (substantial)	0,71 (substantial)
DK	0,66 (substantial)	0,40 (fair)

Table 1: Cohen's kappa (Cohen, 1960) for sentiment annotation. Kappa subscripts indicate the number of categories used. Labels interpreting the degree of agreement from Viera and Garrett (2005).

Inter-coder agreement is generally acceptable, though less so in Danish due to significant positive-neutral disagreement. This is consistent with prior observations about the greater saliency of negativity in political speech (e.g. Young and Soroka, 2012), as that category enjoys greatest agreement in both languages. The discrepancy between the 5-point and 3-point scales also indicate disagreement about intensity, more markedly for Danish.

4. Method

4.1. Word Expansion

Word expansion leverages word embeddings and vector similarity from a large textual corpus to create a bespoke lexicon for genre-specific use. A fuller description of the methodology can be found in Rheault et al. (2016).

Initially, a core set of conceptually neutral sentiment lemmas [1] are expanded using Wordnet

[1] good/bad, love/hate, happy/sad

synonym searches in order to define the relevant constructs: positive and negative valence. This search step was performed manually in DDO (DSL, 2023) for Danish and BulNet (DCL, 2017) for Bulgarian. Words naming political topics, procedures or institutions (e.g. "opposition") were avoided, as their connotations are genre-specific. The 200 most frequent lemmas of the positive and negative synonym lists were retained, based on general language frequency data (Asmussen (n.d.) for Danish and BulTreeBank (n.d.) for Bulgarian). These lists constitute the seed lemmas for the word expansion process.

The lemmatized speech corpus is then embedded using the GloVe algorithm (Pennington et al., 2014), which builds on global co-occcurence data as well as weighted context windows around each word. For disambiguation of homophonous lemmas with different parts of speech, lemmas are concatenated with their PoS-tags before embedding (e.g. "tale-VERB"). Since the ParlaMint corpora are significantly smaller than the British Hansard corpus used by Rheault et al. (2016), the size of the word embeddings was halved to 150 dimensions, and the minimum frequency of a lemma to was set to 5 instances.

The resulting embedding vocabulary is used to construct the final sentiment dictionary. The embeddings for the positive and negative seed lists are separated out of the embedding vocabulary. Similarity scores for the rest of the vocabulary are computed according to the difference between the sum of the cosine similarity of a word to the positive seeds P and to the negative seeds Q:

$$s_i = \sum_{p=1}^{P} \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_p}{\|\boldsymbol{w}_i\|\|\boldsymbol{w}_p\|} - \sum_{q=1}^{Q} \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_q}{\|\boldsymbol{w}_i\|\|\boldsymbol{w}_q\|} \quad (1)$$

Where $\boldsymbol{w}_i$ is the embedding for word i. This similarity construct ensures that high scoring lemmas will be similar to positive seeds while also being dissimilar to negative ones (vice versa for low scoring lemmas) – in order to avoid expanding the seed lists with antonyms, which often appear close to each other in embedding spaces.

The 500 words with the most and least positive scores are used to extend the positive and negative seed lists, excluding numerals and proper nouns. Vector similarity scores are retained as weights meant to correspond to intensity. In practice they also privilege the original seed words which are assigned unit weights (though see Vries, 2022 for a different approach to seed weights). While Rheault et al. (2016) include 1000 lemmas of each polarity, visual inspection revealed that match quality deteriorated much faster due to the smaller size of the ParlaMint corpora, so only 500 lemmas were retained from each group. This yielded a total lexicon size of 1400 lemmas per language, split evenly between positives and negatives.

Speech sentiment is computed by counting and weighting lexicon words (actually lemma-PoS pairs) according to their similarity scores before normalizing by speech length, i.e. a sort of "sentiment density" of speech. To account for negation, words between negation words and punctuation marks are neutralized (i.e. given a default score of 0).

4.2. Other methods

The Polyglot (Chen and Skiena, 2014) and Lexicoder (Proksch et al., 2019) sentiment analyses were implemented based on lexica made available by the authors as replication data. Scoring procedures are also taken from their respective works, which may be consulted for details.

5. Lexicon Evaluation

Sentiment analyses with all three methods were performed for the Bulgarian and Danish gold standards, and results are evaluated based on congruence with with hand-annotated scores, estimated as Spearman correlations due to the ordinal nature of the annotation data. The Polyglot (PG) and LSD lexica are implemented (see Table 2) in both original and lemmatized versions. Inter-lexicon correlations are regular Pearson correlations.

DK	WE_L	LSD	LSD_L	PG	PG_L	$\boldsymbol{AN}$
WE_L	1.0	0.29	0.41	0.34	0.43	**0.47**
LSD	0.29	1.0	0.74	0.40	0.43	**0.35**
LSD_L	0.41	0.74	1.0	0.39	0.49	**0.36**
PG	0.34	0.40	0.39	1.0	0.73	**0.27**
PG_L	0.43	0.43	0.49	0.73	1.0	**0.33**
$\boldsymbol{AN}$	**0.47**	**0.35**	**0.36**	**0.27**	**0.33**	**1.0**

BG	WE_L	LSD	LSD_L	PG	PG_L	$\boldsymbol{AN}$
WE_L	1.0	0.22	0.19	0.20	0.26	**0.36**
LSD	0.22	1.0	0.71	0.47	0.41	**0.36**
LSD_L	0.19	0.71	1.0	0.47	0.41	**0.45**
PG	0.20	0.47	0.43	1.0	0.65	**0.34**
PG_L	0.26	0.41	0.52	0.65	1.0	**0.37**
$\boldsymbol{AN}$	**0.36**	**0.36**	**0.45**	**0.34**	**0.37**	**1.0**

Table 2: Pearson's correlation matrix for Danish (top) and Bulgarian (bottom) lexicon and annotation (AN) sentiment scores. Subscript denotes lemmatization. Boldface marks Spearman correlations to hand-coded scores.

Table 2 indicates that word expansion (WE) correlates most strongly with the hand-annotations in the Danish dataset, while the lemmatized LSD lexicon does best in the Bulgarian dataset. In both cases the difference between best and second-

best is large, but there is little cross-lingual consistency in the relative performance of the lexica.

Lemmatization positively impacts correlations to hand-coded annotations as well as intra-lexicon correlations in both languages. The effects of lemmatization are more pronounced for Bulgarian than Danish, which is consistent with the greater morphological richness of Bulgarian.

A factor that makes these results difficult to interpret is the different scoring methodologies associated with each lexicon. For example, WE scoring involves neutralizing sentiment between negating words and punctuation, while LSD counts are logarithmically transformed (Proksch et al., 2019). In order to test the lexica on an equal footing, we also scored the sample speech excerpts with the simplest counting measure – the difference between positive and negative counts normalized by speech length:

$$s_{speech} = \frac{n_{pos} - n_{neg}}{n_{speech}} \quad (2)$$

Correlation matrices for this new score are presented in Table 3.

DK	WE_L	LSD	LSD_L	PG	PG_L	$\boldsymbol{AN}$
WE_L	1.0	0.26	0.33	0.3	0.31	**0.30**
LSD	0.26	1.0	0.80	0.53	0.49	**0.37**
LSD_L	0.33	0.80	1.0	0.51	0.52	**0.36**
PG	0.30	0.53	0.51	1.0	0.85	**0.32**
PG_L	0.31	0.49	0.52	0.85	1.0	**0.33**
$\boldsymbol{AN}$	**0.30**	**0.37**	**0.36**	**0.32**	**0.33**	**1.0**

BG	WE_L	LSD	LSD_L	PG	PG_L	$\boldsymbol{AN}$
WE_L	1.0	0.32	0.17	0.22	0.20	**0.30**
LSD	0.32	1.0	0.65	0.57	0.47	**0.39**
LSD_L	0.22	0.65	1.0	0.46	0.55	**0.47**
PG	0.17	0.57	0.46	1.0	0.72	**0.34**
PG_L	0.20	0.47	0.55	0.72	1.0	**0.38**
$\boldsymbol{AN}$	**0.30**	**0.39**	**0.47**	**0.34**	**0.38**	**1.0**

Table 3: Pearson's correlation matrix for Danish (top) and Bulgarian (bottom) simplified lexicon and annotation (AN) sentiment scores. Subscript denotes lemmatization. Boldface marks Spearman correlations to hand-coded scores.

For WE, this is essentially its native scoring method without the pre-processing steps of weighting, negation neutralization and PoS-tagging. Across languages, WE performs much worse in this analysis, while the other lexica scores are either unchanged or slightly improved. This indicates that additional processing steps (see Discussion), rather than lexical quality due to domain relevance, account for the good performance of WE in Table 2.

6. Discussion

6.1. Multilingual Performance

Word expansion does not emerge as a language-agnostic method, attaining much greater accuracy for Danish than Bulgarian. In Danish, it outperforms the other methods by a large margin, while it is outperformed by the LSD method in Bulgarian, making its relative performance inconsistent as well. Simplified scoring yields identical but unacceptably low correlations with WE. Other methods have this issue to a lesser degree – for example, lemmatization improves LSD correlations much more strongly for Bulgarian than Danish, creating a disparity between the languages.

In general, the results compare unfavorably with prior work on the same languages. Vries (2022) reports Spearman correlations of 0.46 with hand-coded scores for Danish WE, using a method with some additional optimization. As to the LSD, Proksch et al. (2019) report significantly higher correlations to hand-coded scores for Danish (0.8) and Bulgarian (0.7), reversing the relative performance of the lexicon observed in the current study. This may reflect a difference between native vs. translated text or the quality of the hand-annotations. Only one annotator is used in the study, and the validation evidence in 3.2 suggests a fair amount of uncertainty, particularly for Danish, which is in fact associated with lower correlations across the board. The one exception is the original WE scoring method, which does better in Danish, probably due to the larger corpus size of ParlaMint-DK, which allows for better co-occurrence statistics and semantic embeddings relative to the smaller Bulgarian corpus.

6.2. Pre-processing and Scoring

WE pre-processing also gives Danish another unintended advantage. This study has revealed that lexicon quality and scoring methods interact in complicated ways. For example, the Polyglot lexica performed slightly better with the simplified scoring than with their native scoring method from Chen and Skiena (2014). This highlights the lack of clear theoretical or empirical justification for particular scoring methods in the field. It also indicates that pre-processing procedures appear to affect lexicon performance as robustly than scoring methods. This is mostly clearly seen with WE, where simplified scoring, i.e. native scoring without pre-processing, dramatically affects results.

To further investigate the impact of these steps, we redid the analysis, cumulatively adding WE processing steps – intensity weights, then negation neutralization, and finally PoS-matching (Table 6.2). Each step yielded improvements.

We surmise that weights add information about intensity or uncertainty (as seed words with unit

	DK	BG
lem	0.30	0.30
lem+weight	0.35	0.31
lem+weight+neg	0.41	0.35
lem+weight+neg+pos	0.47	0.36

Table 4: Spearman correlations between WE scores and hand-coded annotations for different scoring methods.

weights are hand-checked) for greater accuracy. Neutralizing negations helps avoid misattribution of sentiment in both languages. Finally, PoS-tagging disambiguates homophones, which results in a big performance increase for Danish relative to Bulgarian. This appears to be due to the large number of polysemic function words in Danish. Since the GloVe word embeddings are created on lemma-PoS pairs, only some function word meanings are included in the expanded lexicon (e.g. ”jo-INTER” but not ”jo-ADV”). Without PoS-tags, many other largely irrelevant function words are counted, muddying the score.

Investigating the degree to which similar pre-processing steps could improve the performance of the other lexica is beyond the scope of this study, but a preliminary experiment with the Bulgarian Polyglot lexicon and negation neutralization surprisingly did not yield better correlation to hand-coded annotations. On the other hand, we have seen that lemmatization of both lexicon and textual data improved scores across the board for LSD and Polyglot dictionaries and is generally advisable, despite not figuring in the original methodologies.

6.3. Theoretical Issues

The problem of relevance raised earlier also leads to questions about the general quality of lexicon entries. Word expansion is based on co-occurrence statistics and vector similarity, which provide semantic information but do not guarantee relevance for sentiment or fully account for syntactic relations. For example, corpus searches on expansion lemmas indicated that a positive lemma in Danish, ”tæt” (en: ”close”), appears to be included solely because of its frequent occurrence in the phrase ”tæt samarbejde” (en: ”close collaboration”). This can lead to a sort of “double-counting”, as words associated with sentiment words in conventional phrases (e.g. modifiers) are inappropriately assigned similar sentiment. On the other hand, WE has the potential to uncover genre-specific connotations, like the verb “констатирам” (en: ”declare”) in the negative Bulgarian list. A corpus search revealed that this verb is in fact usually used in negative contexts, with a connotation of distrust or disavowal – something that a generic lexicon like LSD cannot capture. However, the relative advantages of this are unclear based on correlation results, and ”double-counting” remains a theoretical problem for WE.

6.4. Limitations

This study has certain limitations relating to linguistic coverage and annotation. Evaluating multilingual performance more thoroughly would require a larger sample of languages and parliamentary corpora for generalisability. Multiple annotators per language, as well as larger annotated datasets, would also ensure more reliable estimates. These limitations introduce uncertainty into our evaluation and may account for the various methods’ underperformance relative to prior work.

7. Conclusion

In this study, we implemented the word expansion method from Rheault et al. (2016) to create Bulgarian and Danish sentiment lexica based on the ParlaMint-BG and ParlaMint-DK parliamentary data sets. A gold standard of 300 speeches per language were annotated and used to evaluate the WE lexica in comparison with two other multilingual sentiment lexica: machine-translated versions of the Lexicoder Sentiment Dictionary (Proksch et al., 2019), and Polyglot dictionaries from Chen and Skiena, 2014.

Results are not consistent across languages, where the WE lexicon dominates for Danish, while the lemmatized LSD does best for Bulgarian. Overall, WE performance is either comparable or slightly better than the other lexica, though disambiguating pre-processing (PoS-tagging, negation neutralization and weighting) is necessary to obtain this advantage – other methods do better with simple word counts. Because of the comparable accuracy and more computationally-efficient construction and implementation process, machine translated lexica (in particular, the LSD) appear preferable as scalable, multilingual solutions for parliamentary sentiment analysis. WE may hold more promise for measuring less-studied constructs without existing standard dictionaries (e.g., Hargrave and Blumenau, 2022).

More broadly, this study brings together multiple strands of recent research, underscoring the challenge of achieving language-agnosticism with lexicon methods, as well as the need for comparative empirical evaluation of NLP procedures. Future work will test robustness across a wider swath of languages and further refine and develop current methods for scalable multilingual sentiment analysis.

8. Bibliographical References

Gavin Abercrombie. 2021. *Topic-Centric Sentiment Analysis of Uk Parliamentary Debates*. Ph.D., The University of Manchester (United Kingdom), England. ISBN: 9798515258375.

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, volume 1, pages 2–1.

Sumit Bhatia and Deepak P. 2018. Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Margaret M Bradley and Peter J Lang. 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.

Yanqing Chen and Steven Skiena. 2014. Building Sentiment Lexicons for All Major Languages. pages 383–389.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Seungwoo Han. 2022. Elite polarization in South Korea: evidence from a natural language processing model. *Journal of East Asian Studies*, 22(1):45–75. Publisher: Cambridge University Press.

Lotte Hargrave and Jack Blumenau. 2022. No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK. *British Journal of Political Science*, 52(4):1584–1601.

Borislav Kapukaranov and Preslav Nakov. 2015. Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Sikonja. 2022. Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis. ArXiv:2207.01054 [cs].

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS ONE*, 11(12):e0168843. Publisher: Public Library of Science.

Douglas R Rice and Christopher Zorn. 2021. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, 9(1):20–35.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *acm Transactions on Information Systems (tois)*, 21(4):315–346.

Anthony J Viera and Joanne M Garrett. 2005. Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, 37(5):360–363.

Erik de Vries. 2022. The Sentiment is in the Details: A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media. *Computational Communication Research*, 4(2):424–462. Publisher: Amsterdam University Press.

Lori Young and Stuart Soroka. 2012. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231. Publisher: Routledge.

9. Language Resource References

Asmussen, Jørg. n.d. *The Most Frequent Words in Danish*. Society for Danish Language and Literature (DSL).

BulTreeBank. n.d. *Frequency List*. BulTreeBank Group, Bulgarian Academy of Sciences (BAS).

DCL. 2017. *BulNet (Bulgarian WordNet)*. Department of Computational Linguistics (DCL), Bulgarian Academy of Sciences (BAS).

DSL. 2023. *Den Danske Ordbog (The Danish Dictionary)*. Society for Danish Language and Literature (DSL).

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Fišer, Darja and Pirker, Hannes and others. 2023. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0*. CLARIN ERIC. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1488.