Proceedings of the SALLD-2 Workshop @LREC2020, pages 19–24
Language Resources and Evaluation Conference (LREC 2022), Marseille, 24 June 2022

Evaluating a New Danish Sentiment Resource: the Danish Sentiment Lexicon, DSL

Nina Skovgaard Schneidermann and Bolette Sandford Pedersen
Centre for Language Technology, University of Copenhagen
Emil Holms Kanal 2, DK2300 S
ninasc@hum.ku.dk, bspedersen@hum.ku.dk

Abstract

In this paper, we evaluate a new sentiment lexicon for Danish, the Danish Sentiment Lexicon (DSL), to gain input regarding how to carry out the final adjustments of the lexicon. A feature of the lexicon that differentiates it from other sentiment resources for Danish is that it is linked to a large number of other Danish lexical resources via the DDO lemma and sense inventory and the LLOD via the Danish wordnet, DanNet. We perform our evaluation on four datasets labeled with sentiments. In addition, we compare the lexicon against two existing benchmarks for Danish: the Afinn and the Sentida resources. We observe that DSL performs mostly comparably to the existing resources, but that more fine-grained explorations need to be done in order to fully exploit its possibilities given its linking properties.

Keywords: sentiment analysis, sentiment lexicons, Danish language resources, linguistic linked open data (LLOD)

1. Introduction

As a result of the constantly growing availability of unstructured data, sentiment analysis continues to be of great interest to NLP researchers and industries alike (Liu, 2012). Recent advances in natural language processing have focused on fine-tuning large pre-trained language models such as BERT to the sentiment analysis task, enabling models to automatically extract critical features seen during training (Catelli et al., 2022). Although such approaches yield impressive results, they also tend to be notably data-hungry and may be less flexible for domain-specific tasks (Asghar et al., 2017) and low-resource languages with notable data scarcity (Eskevich et al., 2022).

A complementary method to machine learning approaches is lexicon-based sentiment analysis (Devitt and Ahmad, 2013; Khoo and Johnkhan, 2018): Lexicon or dictionary-based approaches typically make use of a word list containing individual words and matching scores aggregated over a unit of text in a dataset, cf. (Liu, 2012) among others, along with enhancement rules to the scoring mechanism that lifts the model over a simple bag-of-words approach, namely practices for reversal of the sentiment triggered by negation, and for modification of sentiment scores through intensification (Asghar et al., 2017). Although lexicon-based approaches have several limitations in practice, they have the advantage of drawing on information relevant to the domain or the characteristic of the language (Catelli et al., 2022). As such, sentiment word lists can be valuable for low-resource languages, see for instance in Enevoldsen and Hansen (2017) for Danish, where they can either be implemented in a purely rule-based model or as part of a hybrid approach; e.g., sentiment scores from lexica could function as features to a pre-trained language model or a text classifier.

Together with the focus on constructing language-specific sentiment resources, increased attention has also been given in recent years to standardizing and combining such resources, as well as with other linguistic resources as envisaged by the Linguistic Linked Data Community (LLOD), cf. `https://linguistic-lod.org/`. Iglesias and Sánchez-Rada (2021) accounts for the potential of employing standardized formats and tagsets for sentiment resources and making them interoperable and interlinked to an extent where they can be integrated with other NLP datasets and tools and applied together at a large scale.

This paper details the evaluation of a new sentiment lexicon for Danish (Nimb et al., 2022). DSL differs from the existing Danish lexica. It is linked to many other Danish lexical resources via the lemma and sense inventory of the Danish monolingual dictionary (DDO) and the LLOD via the Danish wordnet, DanNet. The evaluation includes a comparison against two existing benchmarks for Danish, namely the Afinn (Nielsen, 2020) and Sentida word lists (Lauridsen et al., 2019), and a more detailed investigation of the DSL resource. Our aim with this paper is twofold: First, we hope to provide input on how to carry out further adjustments to the resource, and secondly, we hope to more generally understand how DSL's linking with other resources contributes to the results. We hypothesize that DSL will perform better than the existing Danish benchmarks due to being more expansive than existing Danish sentiment lists.

The structure of the remaining paper is as follows: In Section 2, we present existing sentiment resources for Danish, and we then go into more detail about the new lexicon and describe its basis on lexicographical principles and linking to other resources. Section 3 describes the pre-processing and implementation steps taken to enhance the lexicon. Section 4 is a comparative evalu-

ation of the three existing word lists, detailing obtained results and more in-depth analyses of the findings. Section 5 discusses the findings in the context of future research, and section 6 contains a summary and conclusions.

2. Relevant Background on Danish Sentiment Lexica

2.1. Existing Danish Sentiment Resources

To our knowledge, Afinn was the first freely available sentiment resource for Danish and is described together with other resources in Nielsen (2020). This sentiment list is a translation and customization of an existing English sentiment lexicon (Nielsen, 2011). The coverage amounts to approx. three thousand lemmas marked with binary polarity values indicate a polarity scale from -5 to $+5$. The resource contains no neutral words.

The more recent and slightly larger sentiment list, Sentida (Lauridsen et al., 2019), contains 5,200 word stems. A background resource for this list was constituted by a list of the 10,000 most frequent Danish words [1], of which all polarity words were selected and neutral words omitted. The list subsumes the words from Afinn and follows the same polarity scaling (-5 to $+5$).

2.2. The New Sentiment Lexicon, DSL, Integrated with other Danish Resources and with the LLOD

The Danish Sentiment Lexicon (Nimb et al., 2022) (henceforth DSL) is a recently published resource based on existing Danish dictionaries, primarily the Danish Thesaurus (Nimb et al., 2014) (henceforth DT). The work is compiled in collaboration between The Danish Society for Language and Literature and The Centre for Language Technology at the University of Copenhagen and funded by The Carlsberg Foundation. The dictionary contains 14,000 lemmas encoded with polarity values from -3 to $+3$, the lowest indicating negative and highest positive values. Less than two-thirds of the words have negative polarity, leaving the rest with positive polarity values. Furthermore, the resource includes morphosyntactic information, namely word classes and a list of word forms for each lemma. This information is not available in either Afinn or Sentida: Afinn contains only word forms, making the number of unique words notably smaller than the actual size of the word list, and Sentida includes words that have been automatically stemmed with the Danish snowball stemmer, which contains some limitations.

The primary purpose of compiling yet another sentiment lexicon for Danish was twofold.

First of all, the development was based on the hypothesis that a higher quality resource could be achieved if it was compiled using monolingual lexicographic methods and resources and not biased by an English source. More specifically, this assumption resulted in DSL being based on the links between groups of words listed in semantic order in a Danish thesaurus, DT (cf. (Nimb et al., 2022) and (Nimb et al., 2014)), and on the corresponding word sense descriptions found in a comprehensive monolingual dictionary, namely The Danish Dictionary, DDO. In short, this meant to identify negative and positive sections in the Thesaurus, extract the words from these sections and combine them with the dictionary information via links. Via the individual thematic areas of DT, the encoders of DSL had available information about synonyms and near-synonyms within a particular topic - also across word classes. The claim is that this background material further eased the calibration of polarity values across word classes and different semantic fields.

Secondly, by being integrated with a collection of Danish lexical resources, the DSL is also being linked to LLOD via the Danish wordnet, DanNet, which has recently been transformed to the Ontolex-Lemon format (Buitelaar et al., 2013). Several RDF polarity relations based on the Marl ontology (`http://www.gsi.upm.es:9080/ontologies/marl/` are defined, and all sentiment data from DSL is made available through the wordnet, with the polarity values percolated down at synset level [2]. This integration with LLOD opens for more extensive use of the sentiment data to be applied in a broader NLP pipeline where other levels of linguistic analysis are compiled and where textual data sets and similar resources for other languages can be taken into account. Combining cross-lingual data with purely monolingually defined data in DSL could potentially improve the usability of the resource.

3. Experiments

Our experiments consisted of implementing a model and evaluating it against existing benchmarks on four manually annotated sentiment datasets, all of which were made publically available through the DaNLP repository (Pauli et al., 2021). The datasets are as follows:

- EuroparlSentiment1. It consists of 184 sentences from sections of the Danish part of the Europarl Corpus (Koehn, 2005). The sentences are manually annotated with polarity scores between -3 and 3 by Nielsen (2020)

[1] The list was achieved from The Danish Society for Language and Literature: `https://korpus.dsl.dk/resources/details/freq-lemmas.html`

[2] Note that DSL was encoded with a basis in the DDO and therefore originally encoded at sense level. Lemmas that had several senses with diverging polarity were carefully studied. Half of these were rejected due to ambiguity (e.g., frelst ('saved'), sej ('tough'), skarp ('sharp'), overlegen ('superior') and glat ('smooth'). The other half was kept in the lexicon since it was estimated that the polarity sense was by far the most frequent sense of the lemma (Nimb et al., 2022)

- LCCsentiment: Consists of 499 sentences from sections of the Leipzig Corpora Collection (henceforth LCC) (Biemann et al., 2007), likewise annotated by Nielsen (2020) in the same way as Europarl1.
- EuroparlSentiment2. It consists of additional 957 sentences from Europarl annotated by the Alexandra Institute (Pauli et al., 2021). The dataset contains both subjectivity and polarity scores, although only polarity values are measured in this instance. Polarity values are annotated as 'negative', 'positive', or 'neutral'.
- TwitterSentiment. It consists of 1413 tweets annotated by the Alexandra Institute with negative, positive, and neutral polarity labels (Pauli et al., 2021).

Our model implementation consists of a search function that matches the lexicon against the dataset to be searched and a scoring function that aggregates the sentiment scores over every sentence in the dataset. The following sections will describe the pre-processing steps on the data, along with additional rules and enhancements implemented to increase the scoring accuracy. Finally, the measures of evaluation are briefly discussed.

3.1. Pre-processing

Before the search is conducted, the data is tokenized and POS-tagged using DaCy as a pre-processing step. This Danish pre-processing framework has achieved state-of-the-art performance on POS-tagging and named entity recognition (Enevoldsen et al., 2021). The data is then lemmatized with tokens and POS-tags as inputs using Lemmy [3], a python-based Danish lemmatizer trained on the Danish full-form list from DDO and the Universal Dependencies converted from the Danish Dependency Treebank (DDT) (Johannsen et al., 2015). This step was taken to utilize the morphosyntactic information available in DSL, word classes, and homographs, to disambiguate words in the data when possible (see 3.2.). It, therefore, provides an example of how linguistically linked data has been employed to increase the flexibility of our model.

Furthermore, a stopword filter was applied to the tokens to decrease noise during scoring. We made a manual assessment of the subset of the 219 words in the original stopword list, which would be useful for sentiment scoring, consisting of adverbial modifiers 3.2. along with six lemmas, primarily adverbs, which were present in DSL:

- 'Måske' ('maybe'): -1
- 'Nemlig' ('in fact'): 2
- 'Skulle' ('have to', 'should'): -1
- 'Alene' ('alone'): -1
- 'God' ('good'): 3
- 'Allerede' ('already'): 1

Conversely, the stopword list also included instances of words that were present in DSL but which have ambiguity issues that can currently not be solved: An example of this is 'du,' which in Danish is ambiguous between the 2nd person singular pronoun and the infinitive form of the verb 'to function.' Since there is currently no implementation to effectively deal with cases where a sentiment-bearing word is ambiguous with a frequent, non-sentiment-bearing word, the presence of the lemma would contribute to more noise than useful information during the search and was therefore filtered out.

3.2. Model enhancements

Our enhancement rules consist of two components: Disambiguation rules in cases of homographs and sentiment-modifying rules in the presence of negators and intensifiers.

DSL is the only one of the three lists containing homographs, i.e., duplicate sentiment-bearing lemmas with different meanings and sense-level information and parts of speech from DDO. This makes it possible to implement simple disambiguation procedures in cases where sentiment-bearing homographs were found during matching: For this purpose, we map the part-of-speech information in DSL to the automatic POS tags generated by DaCy and match them against the data. If a matching POS tag is found for a given ambiguous lemma, the model chooses the corresponding sentiment score and drops the remaining ones. Otherwise, it takes the 1st sense of the word in DDO to be the correct one, as this is typically also the most frequent.[4]

Additionally, a series of heuristics (Lauridsen et al., 2019) for dealing with sentiment-modifying elements were applied: Sentiment scores are reversed in the presence of negation if a sentiment-bearing word exists within the scope of -1 to $+3$ positions from the negation trigger. Other elements that have been found to increase or reduce sentiment include intensifying adverbial modifiers ('very,' èxtremely,' 'slightly' etc.), the conjunct 'but,' which could be said to weaken the statement expressed by the preceding clause, and exclamation marks and all-caps, which both increase the score (Dragut and Fellbaum, 2014; Asghar et al., 2017). We applied a dictionary of adverbial modifiers and their corresponding values, which were initially described for English by (Dragut and Fellbaum, 2014) and adjusted for Danish by Lauridsen et al. (2019). The values are multiplied with the total sentiment score if an adverbial modifier is proceeded by a sentiment-bearing word.

[3] See `https://github.com/sorenlind/lemmy`

[4] It should be noted that the SpaCy POS-tags are not in one-to-one correspondence with the word classes in DDO, which may contribute to some inaccuracies.

3.3. Evaluation

The three resources were evaluated using two different metrics: First, we calculated the Pearson rank correlation coefficients between a given lexicon and the human-annotated sentiment scores for each dataset. Secondly, we divided the scores outputted by DSL for each dataset into negative, neutral, and positive classes following the procedure for existing DaNLP sentiment benchmarks [5]. This enabled a more direct evaluation of the datasets annotated with 3-way polarity. To account for the imbalance towards words with negative polarity in DSL (62 %), we trained a logistic regression classifier on 990 examples of the TwitterSentiment dataset and adjusted the optimal threshold value, which is given by $max(tpr - fpr)$, where tpr denotes the true positive rate and fpr the false positive rate (Flach, 2010). In accordance with the procedure described by (Pauli et al., 2021), neutral class is taken to be on a continuum rather than a discrete value. Thus, we set the threshold to 0.37 and take scores between -0.37 and 0.37 to belong to the neutral class, scores above 0.37 to be positive, and scores below -0.37 to be negative.

4. Results and analyses

Table 1 provides an overview of the results of the comparative evaluation on each dataset. Table 2 reports on the recall, precision, and micro F1-score for the negative, neutral, and positive classes on DSL.

Dataset	Lexicon	Corr.	Acc.	Avg. F1	Wgt. F1
Europarl1	DSL	**0.703**	**0.685**	0.675	0.676
	Sentida	0.671	0.669	0.651	0.657
	Afinn	0.634	**0.685**	**0.676**	**0.681**
LCC	DSL	0.512	0.639	0.593	0.639
	Sentida	0.526	0.581	0.548	0.579
	Afinn	0.516	0.655	0.606	0.652
Europarl2	DSL	0.459	0.543	0.533	0.541
	Sentida	0.473	0.533	0.514	0.527
	Afinn	0.413	0.557	0.547	0.560
Twitter-Sentiment	DSL	0.387	0.462	0.448	0.470
	Sentida	0.396	0.423	0.416	0.424
	Afinn	0.334	0.478	0.46	0.485

Table 1: Comparative evaluation of Danish sentiment resources.

4.1. Analyses

Overall, we can observe that DSL appears to perform comparably to the existing word lists, with the most significant improvement being a Pearson correlation of 0.70 with EuroparlSentiment1 against 0.66 and 0.63 on Sentida and Afinn, respectively. However, in most cases, DSL does not perform notably better than either word list; in fact, it does not exceed Afinn on classification of tweets, which may be due to the fact that Afinn contains several more colloquial phrasings specific to the domain of social media (Nielsen, 2011). We also observe notable differences between the performances for the evaluated datasets, part of which could be due to significant differences in class distributions: The neutral class in EuroparlSentiment2 comprises nearly half of the samples, whereas only about a fifth of the TwitterSentiment samples are marked as neutral. Furthermore, the overall score appears to decrease with increasing sample sizes, suggesting that the relatively high scores on EuroparlSentiment1 may be a product of few example sentences.

Dataset	Class	Precision	Recall	F1
Europarl1	Negative	**0.781**	0.472	0.588
	Neutral	0.679	0.679	0.679
	Positive	0.634	**0.9**	**0.744**
LCC	Negative	0.474	0.383	0.424
	Neutral	0.727	0.672	0.698
	Positive	0.581	0.758	0.658
Europarl2	Negative	0.593	0.414	0.488
	Neutral	0.654	0.484	0.556
	Positive	0.43	0.768	0.551
Twitter-Sentiment	Negative	0.744	0.411	0.529
	Neutral	0.314	0.359	0.335
	Positive	0.372	0.68	0.481

Table 2: Metrics for each class in DSL on evaluated datasets.

By examining the errors manually, however, we can learn a lot about what may contribute to the relatively minor differences between DSL and the other word lists, in spite of our hypothesis that its expansiveness would yield more reliable sentiment scores: Namely, an inspection of the 1000 most frequent words over all the datasets reveals that the proportion of matched words in DSL only comprises 260 of the 14000 lemmas, of which 225 intersect with Sentida [6]. This may indicate that although the DSL resource may be more expansive in a linguistic sense, it may not make a substantial difference in practice within the relatively conventional domain of politics, news, and social media. In fact, inspecting some of the instances of falsely rated sentiments suggests that DSL may even be too exhaustive in its attribution of sentiment: Namely, words such as 'skulle' ('should, have to') and 'sidste' ('last') are given a sentiment score of -1 and 1, respectively, although examples such as, 'parlamentet skal træffe en beslutning' ('the parlament need to make a decision'), and 'det er deres sidste chance' ('it is their last chance') suggests contexts where a more neutral attribution may be warranted. Other examples of debatable sentiments

[5] `https://github.com/alexandrainst/danlp` (Pauli et al., 2021)

[6] Note that stemming was performed on the DSL lemmas to determine this

are adverbials such as 'måske' ('maybe'), 'allerede' ('already'), and àlligevel' ('still'), which may be better suited as modifying the sentiment of a given sentence than being given their own values. A final point of observation is that DSL is the only one of the three lists containing multiple word senses, which, as seen in 3.1., can cause problems for a rudimentary analysis.

5. Discussion

The results displayed in 4. strongly suggest that a rudimentary evaluation may not be sufficient to uncover the assumed benefits of a more exhaustive sentiment lexicon, particularly with respect to its linked data properties. This is primarily because models that fully utilize the lexicon's linking to DanNet have not yet been implemented given that the resource is relatively recent. As a future line of research, it may be advantageous to investigate the effectiveness of DSL for domain-specific ontology-based approaches to sentiment analysis. The interoperability of the DSL with sense-level information from DanNet and RDF polarity relations based on the MARL ontology would potentially make the graded polarity scores valuable as linguistic features in an aspect-based sentiment model. Developing formal representations of how concepts are related within a given subdomain has been shown to improve both accuracy and flexibility of sentiment models, since it enables a fine-grained overview of public sentiment towards specific topics (García-Díaz et al., 2020). Generally, understanding how DSL may benefit domain-specific flexibility is recommended.

6. Conclusion

This paper has detailed the efforts to evaluate the new Danish Sentiment Lexicon, DSL, which is being linked to the LLOD. We experimented on 4 labelled datasets and performed rudimentary pre-processing of the data, and employed basic rules designed to lift the model slightly over a bag-of-words approach, as well as to take advantage of sense-level information provided by the lexicon. While our rudimentary analyses were not able to verify the effectiveness of DSL over other lexica, it was confirmed that DSL performs comparably with existing Danish word lists in a basic setting. However, in order to fully exploit the possibilities provided by the linking of DSL with other resources, more complex implementations need to be made, an example of which is employing the lexicon for more fine-grained ontology-based sentiment models within specific domains.

Bibliographical References

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., and Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PloS one*, 12(2):e0171649.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, 2007.

Buitelaar, P., Arcan, M., Iglesias, C. A., Sánchez-Rada, J. F., and Strapparava, C. (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 1–8.

Catelli, R., Pelosi, S., and Esposito, M. (2022). Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374.

Devitt, A. and Ahmad, K. (2013). Is there a language of sentiment? an analysis of lexical resources for sentiment analysis. *Language resources and evaluation*, 47(2):475–511.

Dragut, E. and Fellbaum, C. (2014). The role of adverbs in sentiment analysis. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 38–41.

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

Enevoldsen, K., Hansen, L., and Nielbo, K. (2021). Dacy: A unified framework for danish nlp. *arXiv preprint arXiv:2107.05295*.

Eskevich, M., de Jong, F., Giagkou, R. M., and Hajič, J. (2022). Project european language equality (ele) grant agreement no. lc-01641480–101018166 ele coordinator prof. dr. andy way (dcu) co-coordinator prof. dr. georg rehm (dfki) start date, duration 01-01-2021, 18 months.

Flach, P. A., (2010). *ROC Analysis*, pages 869–875. Springer US, Boston, MA.

García-Díaz, J. A., Cánovas-García, M., and Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:641–657.

Iglesias, C. A. and Sánchez-Rada, J. F. (2021). Sentiment analysis meets linguistic linked data. *Proceedings from SALLD-1 2021*, 2021.

Johannsen, A., Alonso, H. M., and Plank, B. (2015). Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Lauridsen, G. A., Dalsgaard, J. A., and Svendsen, L. K. B. (2019). Sentida: A new tool for sentiment analysis in danish. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(1):38–53.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Making Sense of Microposts (MSM2011), 93-98*.

Nielsen, F. A. (2020). Danish resources. Technical report, Danish Technical University, Lyngby.

Nimb, S., Lorentzen, H., Theilgaard, L., and Troelsgård, T. (2014). *Den danske begrebsordbog*. Det Danske Sprog-og Litteraturselskab.

Nimb, S., Olsen, S., Pedersen, B. S., and Troelsgaard, T. (2022). A thesaurus-based sentiment lexicon for danish: The danish sentiment lexicon. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Marseille, France, May. European Language Resource Association (ELRA).

Pauli, A. B., Barrett, M., Lacroix, O., and Hvingelby, R. (2021). Danlp: An open-source toolkit for danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466.