EMMA: Danish Natural-Language Processing of Emotion in Text

The new State-of-the-Art in Danish Sentiment Analysis and a Multidimensional Emotional Sentiment Validation Dataset

Esben Kran (201909190@post.au.dk)

Søren Orm (201907685@post.au.dk)

University of Aarhus, Department of Cognitive Science

Poetry is when an emotion has found its thought and the thought has found words. - Robert Frost

1 Abstract

In this paper, we report on the current state of Danish text sentiment analysis (SA) and the process of creating the new state-of-the-art Danish SA tool by updating the current SENTIDA. Our SentidaV2 shows significant improvement in classifying sentiment in text compared to the original SENTIDA tool ($p < 0.01$) in three different validation datasets. We also present a new domain general validation dataset called Emma (Emotional Multi-dimensional Analysis) that exceeds current validation methods in both complexity and quality, created with a proprietary adaptive tool for supervised machine learning data collection utilizing a citizen science approach. Emma consists of sentences scored in a four-dimensional emotional circumplex space by 30 coders with a wide demographic representation that allows for future multidimensional, fine-grained machine learning-based Danish SA. We also publish Emma and SentidaV2 optimized for Python on GitHub.

Keywords: Sentiment Analysis, Danish NLP, Computational Linguistics, Dataset

2 Introduction

In order to discover new patterns, trends, and make predictions in the ever-growing amount of information available to us in the digital age, we need new tools for analysing this information. A considerable part of the information is encrypted in a complex and notoriously difficult to decipher format called 'language', and each language needs its own set of tools for computational analysis. We currently have tools that can assess the sentiment in text for Danish, however, these tools can not only be optimized to better match international standards, but can also be expanded so they are able to assess the complex range of emotions that people are able to experience and express in written language. This paper seeks to further develop the tools available in the field of Danish computational linguistics.

How can the current state-of-the-art in Danish sentiment analysis be improved? This paper proposes that the sentiment scoring accuracy can be increased through higher syntactical and semantic awareness in SENTIDA and that by creating a new, coded validation dataset for SA, the field can approach practical multi-dimensional Danish emotional SA. This paper introduces SentidaV2 and Emma. SentidaV2 build upon existing SA models with further context awareness to increase accuracy. Emma is a dataset of sentences rated on four emotional dimensions of valence (positivity), intensity, controllability, and utility.

3 Sentiment Analysis

SA is a part of applied linguistics and attempts to quantify the positivity of natural language, especially on the internet or in large corpora of

texts like newspaper databases. An example of a use case is extracting the sentiment score of the sentence *'Lad der blive fred'* (*Let there be peace*) that gets a sentiment score of 2 with SentidaV2 on a scale of -5 to 5 which indicates a positive valence above 0.

In the field of sentiment analysis (SA), sentiment is the presence of negative or positive charge in a word, sentence, or larger piece of text (B. Liu, 2012; Mäntylä et al., 2018). Approaches to analysing sentiment differ widely in complexity from a bag-of-words approach (BoW), where the sentiment of the input is determined by matching words to a sentiment lexicon; to aspect-aware neural network (NN)-based approaches with advanced context awareness influencing the same word's sentiment score based on its surroundings (Hoang et al., 2019; N. Liu et al., 2019); and multidimensional valence-arousal (VA) SA using novel combinations of NN techniques (Maas et al., 2012; Wang et al., 2016). The current standard in Danish (Lauridsen et al., 2019; Nielsen, 2017) builds on a BoW approach with limited syntactic awareness, and with a unidimensional scale of valence based on the circumplex model of affect (Russell, 1980). This unidimensional scale varies from -5 to 5 and indicates the level of negative vs. positive emotion associated with specific words.

These are generally only on one or two dimensions despite the circumplex models developed in emotional analysis with newer emotional models like the three-dimensional measurement of emotions with "dominance" (Bradley & Lang, 1999; Osgood et al., 1957) and the modern 4D representation of emotion that challenges the circumplex model and adds both "controllability" and "utility" to the VA scales (Trnka et al., 2016). The next step in emotional SA should incorporate these.

3.1.1 Bag of Words (BoW) approach

Current SA tools use a semi-BoW approach to sentiment analysis, where a word is associated with the relevant sentiment score (Table 1) (Hutto & Gilbert, 2014). The aggregate sentiment scores of the words in the text are then used as an indication of how positive the text is.

Table 1 - Example of lexicon words with sentiment score

'Accept' (*acceptance*)	1.5
'Advarsel' (*warning*)	-2

3.1.2 Neural network approaches

Contemporary neural network models such as word2vec, FastText, and BERT (Bojanowski et al., 2017; Devlin et al., 2019; Goldberg & Levy, 2014; Grave et al., 2017; Howard & Ruder, 2018; Joulin et al., 2016; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Peters et al., 2018) represent the state-of-the-art NNs for sentiment analysis. Further developments of these architectures have also given rise to aspect extraction methodologies to perform aspect-based sentiment analysis (ABSA) (Hoang et al., 2019; N. Liu et al., 2019; Rana & Cheah, 2016; Shafie et al., 2018), where a single word is assigned different sentiment scores based on the context.

For dimensional valence-arousal sentiment analysis, the current state-of-the-art uses combinations of convolutional NN and long short-term memory algorithms (Wang et al., 2016) with circumplex VA space-based training datasets. With a large enough dataset, the same can be achieved on four-dimensional emotional analysis in Danish texts.

4 Current Status

4.1 Current Danish lexical SA

The first SA tool for Danish, AFINN, stemmed from an interest in Twitter sentiment analysis (Nielsen, 2011) and was developed from machine translations of English sentiment lexicons (Nielsen, 2019). It currently consists of 3,552 rated words in Danish and 96 rated emoticons (Nielsen, 2015/2019).

SENTIDA is a lexicon consisting of the 5,263 most-used Danish sentiment-carrying lemmas (Lauridsen et al., 2019). These words were separately rated for positive vs. negative sentiment by the three authors, and the mean rating is used as sentiment score (Lauridsen et al., 2019). Words that did not overlap between AFINN and SENTIDA were copied from AFINN and re-rated by the SENTIDA team.

4.2 Difficulties in SA

4.2.1 Problems in BoW

Only using the BoW approach has four main problems for SA on texts:

1. Ignores syntactic relations between words. Relationships like verb-noun structure are ignored and generalized which limits the use case accuracy.
2. Ignores intensity modifications of words as these depend on syntactical relation like proximity, e.g., *not* or adverbs have no effect on the subsequent words.
3. Does not reflect human sentiment perception. Humans use pattern recognition and context knowledge to read the sentiment of texts. This is approachable with state-of-the-art methods and might be implementable with neural network (NN) compositions (Wang et al., 2016).
4. No difference in homonyms or contexts as a word can only appear once in the same form.

Alleviations for the BoW approach are introduced in AFINN, SENTIDA, and VADER to differing degrees such as adverbial intensification modifiers, exclamation mark multiplier, and negations. In SentidaV2 these are improved compared to SENTIDA.

4.2.2 Problems in NN

Even though NN-based architectures have high accuracy, their processing speed is very slow, and a lot of good quality data is required to train them which creates problems with the workload of the dataset development as well as the time for processing the text itself during text analysis.

4.3 Potential of quantitative dimensional measures of emotion for SA

Beyond Ekman's basic emotions of anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992), the focus in Emma is on the dimensional quantitative models of emotion such as the circumplex model of affect (Russell, 1980), the PANA model (Watson & Tellegen, 1985), Plutchik's model (Plutchik, 2001), PAD (Mehrabian, 1980), and the hypercubic semantic emotion space (HSES) (Trnka et al., 2016).

The models used for Emma are the HSES and the PAD that consists of four dimensions of emotions: 1. Valence, 2. Intensity, 3. Controllability, 4. Utility (Mehrabian, 1980; Trnka et al., 2016) that also arise from earlier models (Osgood et al., 1957) and introduces new capabilities to modern SA (Poria et al., 2018; Wang et al., 2016)

The potential of these models in SA is the ability to identify different emotions in text beyond the normal one-dimensional positivity or valence scale. Using all four dimensions of the HSES model might not be necessary but by ensuring the maximum amount of empirically based dimensions, the model can always be reduced to a more practical complexity level. As the emotions are identified in the four-dimensional space of the HSES model, by fine-tuning an already existing NLP model using Emma will make Danish emotional SA possible.

5 Improvement process

5.1 SentidaV2

The SentidaV2 Python tool, like AFINN and SENTIDA takes a sentence and splits it into individual words and saves the order of the words. It matches the individual words in the sentence with a list of annotated words. If a given word is not annotated, it receives a rating of 0. The words were annotated by the teams behind SENTIDA and AFINN, 4 people in total, on a scale from –5 to +5, with –5 corresponding with a very negatively charged word and +5 with a very positively charged.

In this paper, we improve on the available tool, by expanding on SENTIDA by adding several intensity modifiers such as: Adding *'ikke'* (*not*) synonyms and abbreviations, e.g. *'ik'* and *'ikk'* (*not*), *'aldrig'* (*never*), and *'ingen'* (*none*):

"Det er **aldrig** (-1 X →) godt (+2,3)."

"That is **never** (-1 X →) good (+2,3)."

Updating *'ikke'* (*not*) in questions – in Danish, the usage of not in a question doesn't negate the sentence:

"Er det ikke (~~-1 X →~~) forkert**?**"

"Isn't (~~-1 X →~~) that wrong**?**"

Adding *'men'* (*but*) intensity updates – in a sentence containing the word *'but'*, the part of the sentence after *'but'* carries more sentimental charge than the part before *'but'*. The English sentiment analysis program VADER uses the factors 0.5 for the part of the sentence before *'but'* and 1.5 for the part after *'but'* (Hutto, 2014/2019). We have adopted these values:

> "Maden (+0.3) var god (+2.3), (← X 0.5) **men** (1.5 X →) serviceringen (+0.3) var elendig (-4.3)."
>
> "The food (+0.3) was good (+2.3), (← X 0.5) **but** (1.5 X →) the service (+0.3) was horrendous (-4.3)."

Adding intensity modulation if an exclamation mark is detected in a sentence. For each exclamation mark (EM) detected in a sentence, the sentiment of the sentence is multiplied by 1.291 for the first, 1.215 for the second, and 1.208 for the third. If more than three EMs are detected, the additional EMs are ignored, and the count of EMs is set to 3. These values are the same used in VADER (Hutto, 2014/2019):

> "Det er så sejt (+3.6)**!** (← X 1.291)"
>
> "It is so cool (+3.6)**!** (← X 1.291)"

Adding uppercase intensity modification. If a word is written in all capital letters, the sentiment of that word is multiplied by 1.733. This values is the same used in VADER (Hutto, 2014/2019):

> "DET ER SÅ SEJT (+3.6). (← X 1.733)"
>
> "IT IS SO COOL (+3.6). (← X 1.733)"

We also expand on SENTIDA, which is written in a programing language called R usually used for statistical analysis by translating it to another programming language called Python because Python is more supported in the NLP field. Writing SENTIDA in Python thus eases the process of incorporating improvements made for other languages and using SentidaV2 with other NLP tools.

5.2 Emma

We also introduce a new validation dataset, Emma. Until now, Danish SA has been validated on TrustPilot reviews, trying to guess whether a review is positive (having 4 or 5 stars) or negative (having 1 or 2 stars). TrustPilot reviews are used to validate SA programs mainly because it's easy to acquire a large set of rated sentences. However, using TrustPilot reviews has its problems: Spelling mistakes occur often and the SA program will not be able to recognize the words; rating mistakes happen, causing the validation to lose accuracy; and there is no clear etiquette for how to write or rate reviews, which is a problem because some might in a 4-star review write why the product got 4-stars, while others might write why it didn't get 5-stars. The same experience with a product might cause two different consumers to write similar reviews while rating the product differently, too.

As opposed to the TrustPilot reviews, Emma consists of 352 sentences rated on 4 dimensions by 30 raters, who all received the same instructions in how to rate the sentence. The raters were found using citizen science, to ensure a broad demographic representation of Danes. Citizen science also contributes to motivation for the coders as it attempts to allude to their sense of assisting in a scientific endeavor which has been shown to increase engagement (Heck et al., 2018; Pedersen et al., 2017).

Dimension	Valence	Intensity	Controllability	Utility
Danish scheme	Meget negativ følelse	Meget beroligende	Meget ukontrollabelt	Meget skadeligt
	Meget positiv følelse	Meget ophidsende	Meget kontrollabelt	Meget gavnligt
Translation	Very negative feeling	Very calming	Very uncontrollable	Very harmful
	Very positive feeling	Very arousing	Very controllable	Very beneficial

Table 2 - Coding scheme for the citizen science raters

5.3 Emma data collection program

The program takes the form of a form that is the user interface for the coders and includes the coding scheme along with a structure optimized for ease of action, so it does not seem intimidating for the citizen science (CS) coders. The coding scheme is structured into the four different emotional dimensions as seen in Table 2.

Coders are then asked to rate 20 sentences on each dimension from -5 to 5 with neutral = 0 after an introduction to the scheme, Citizen Science, the project itself, the mechanics of the survey, and a top five leaderboard for the gamification incentive functionality is displayed.

After the encoding process, the respondents are asked to provide information about their demographic such as ID, region, age, educational level, and occupation. If the ID has been used before, the points for that ID are updated after submission of the form to reflect their position on the leaderboard.

5.4 Coders

To ensure a wide demographic representation in the text annotation process, the annotation software was distributed in social networks consisting of a wide range of Danish citizens from different regions, educational levels, occupations, and ages. This was done in response to the fact that available SA tools do not have a large representation of demographic variety (Lauridsen et al., 2019; Nielsen, 2017).

The demographic variety of our 30 coders spans all levels of society representing different occupational situations (jobless to employers) and job titles, educational backgrounds (middle school to Ph.D.), age ranges (17-65 years), and location (all five Danish regions are represented).

5.5 Intercoder agreement

Measurement of the intercoder reliability was performed using a one-way pairwise interclass correlation coefficient (ICC) test average (Fleiss & Cohen, 1973; Weir, 2005) with the IRR package (Gamer et al., 2019) in R (R Core Team, 2013). The test assumes random intercepts for each rater, as different raters vary in their baseline subjective standards for text ratings. The ICC of the raters in Emma was 0.743, which is defined as *good* by Cicchetti (1994) and as *moderate* by Koo & Li (2016). For comparison, the interrater reliability of the SENTIDA dataset has a Krippendorff's alpha value of 0.667 (Lauridsen et al., 2019), which is an acceptable value, with >0.8 defined as a good measure (Krippendorff, 2004).

5.6 Ratings

The ratings returned from the validation dataset Emma defines each sentence as a dot in the four-dimensional space representing the hypercubic definition of emotional space (Trnka et al., 2016) (Figure 1). The coordinates for each sentence correspond to the mean values of ratings for that sentence given by the coders.

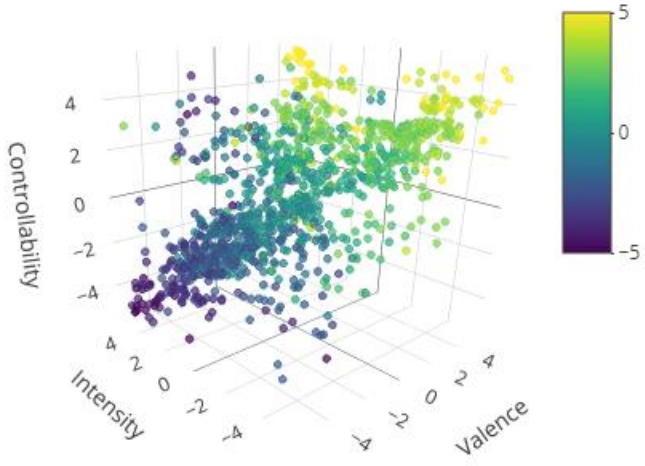

Figure 1 - Sentences and their ratings in the four dimensions (color: utility)

6 Test results

We used the three validation sets TP (used to validate SENTIDA (Lauridsen et al., 2019)), TP2, and Emma. TP and TP2 consists of respectively 7019 and 7015 reviews from the website Trust Pilot along with the number of stars the person writing the review gave the company. TP and TP2 only contain reviews that got 1, 2, 4, and 5 stars. The reviews in TP are lowercase without punctuation, and the reviews in TP2 have their original casing and punctuation. TP2 is used to ensure that all the improvements made in SentidaV2 are utilized. Emma is a validation set consisting of 352 rated sentences. The corpus of sentences was

	TP	TP: 95% CI	TP2	TP2: 95% CI	EM	EM: 95% CI
SentidaV2	0.8063	0.7891 to 0.8226	0.8183	0.7999 to 0.8365	0.6957	0.5889 to 0.7889
SENTIDA	0.8052	0.788 to 0.8215	0.7817	0.7623 to 0.8015	0.6748	0.567 to 0.7704
AFINN	0.7497	0.731 to 0.7676	0.7494	0.7284 to 0.7695	0.6371	0.5281 to 0.7366

Table 3 - Results from accuracy tests (TP: TrustPilot, TP2: TrustPilot 2, EM: Emma)

carefully selected in order to represent a wide variety of sentiments and complexity. The sentences were annotated by CS volunteers.

6.1 Validation process

To assess how good SentidaV2 is at classifying sentiment in sentences and how well SentidaV2 performs compared to other Danish SA-programs, we processed the reviews in TP, TP2, and Emma with SentidaV2, Sentida, and AFINN to convert the sentences from text to a sentiment score. For each program, a logistic regression that predicts whether the reviews belong to the 1-star and 2-star group / negative sentiment or to the 4-star and 5-star group / positive sentiment is developed.

Based on these scores, the models predict the scores for TP and TP2. For Emma the sentences are classified as positive if the sentiment score is above 0 and negative if the sentiment is below 0. Again, a logistic regression is developed for each program to predict whether the sentence belongs to the positive or negative category. In order to test the accuracy of the different programs, we use 75% of the reviews as a training set for the logistic regression and set up a confusion matrix on the remaining 25% of the reviews as earlier papers have done (Lauridsen et al., 2019).

This is repeated 1,000 times using a custom function with different training and test sets for each iteration and the average accuracy and average 95% confidence interval extracted for each dataset (Table 3). A t-test between the average performances of SentidaV2 and the current best tool for Danish SA, SENTIDA (Lauridsen et al., 2019) is then set up to see if there is a significant difference between the performances of the two programs.

6.2 Predicting Sentiment

In Table 3 the average accuracies and the average 95% confidence intervals of the three SA-programs, AFINN, SENTIDA, and SentidaV2, on the three validation sets are summarised. On average, SentidaV2 was found to have a significantly higher accuracy at binarily classifying the sentiment of the sentences in TP (M = 0.8063, SD = 0.007), in TP2 (M = 0.8183, SD = 0.008), and in Emma (M = 0.6957, SD = 0.042), compared to the accuracy of SENTIDA for the sentences in TP (M = 0.8052, SD = 0.007), in TP2 (M = 0.7817, SD = 0.008), and in Emma (M = 0.6748, SD = 0.043). The t-values are $t_{TP}(1997.2) = 3.2837$, $t_{TP2}(1988.6) = 98.488$, and $t_{Emma}(1997.3) = 10.995$, and the differences are significant for all datasets ($p_{TP} = 0.001$, $p_{TP2} < 2.2e\text{-}16$, and $p_{Emma} < 2.2e\text{-}16$).

7 Does SentidaV2 and Emma improve Danish SA?

Using the validation set TP poses a few problems. All the words are lower case, meaning no effect from capitalization. There is no punctuation in the reviews, meaning no punctuation effects like *'?'* and *'!!'*. Additionally, each review consists of multiple sentences without punctuation which makes it hard to split them up. This is a problem in sentences containing '*men'* (*but*) because the sentiment modulation is only meant to be applied on sentence to sentence basis. We see a significant difference between the performance of SENTIDA and SentidaV2, but the difference is miniscule.

The reviews in the TP2 validation set, however, have both the original punctuation and casing. This means that a wider range of the

improvements implemented from SENTIDA to SentidaV2 can be tested, and, as expected, there is a positive difference in the accuracy of SentidaV2 on TP2 compared with TP. The difference in performance is especially prominent when compared to the performance of SENTIDA on TP2. There is a larger significant difference between the accuracy of SENTIDA and SentidaV2.

The same pattern is observed for the Emma sentences; a quite substantial and significant difference was found between the accuracy of SENTIDA and SentidaV2. As it has been shown before (Lauridsen et al., 2019), AFINN is outperformed by SENTIDA. This is consistent with our findings.

Regarding Emma, none of the SA programs perform as well on Emma as they do on the two TrustPilot validation sets. A few reasons might be the cause for this difference in performance. First, the TrustPilot reviews are filtered so only the polarities, 1-, 2-, 4-, and 5-star reviews are included, whereas the middle 20% of the sentences of Emma are not filtered out. The sentiment scores of these middle cases are more minute and subtle and thus more difficult to correctly classify and some might not even be suited for binary classification, e.g. if the sentiment is neutral. Secondly, the sentences in Emma display a more complex and context dependent usage of language not necessarily having an obvious positive or negative sentiment as opposed to the TrustPilot reviews, where the context is given, i.e. people write about their experiences with a product often explicitly positively or negatively. This can be said to better reflect real-world situations.

The biggest limitation of Emma is its size. In order to ensure optimal validity of Emma, the validation set needs a larger corpus of annotated sentences and a larger number of annotators per sentence. Increasing the number of sentences will ensure that the SA programs validated with Emma will be tested on a wider variety of the Danish language. Increasing the number of ratings per sentence will ensure higher validity of the ratings the sentences have received.

Emma reflects real-world scenarios better but is missing the large amount of data available from e.g. TrustPilot and presents a larger challenge for the SA tools through its complexity than TP and TP2.

8 Future research

Danish sentiment analysis is still far from perfect and needs further development.

As examples, it currently relies on the less than optimal stemming tool 'SnowballC'. The great advantage of the tool is that it expands the number of rated words from 5263 to an estimated 30.000 words (Lauridsen et al., 2019) by reducing different inflections of a word to its root. This also improves the speed of the program. This comes at a price however as not all roots have the same sentiment as their inflections and some words become grossly mis-rated – e.g. the word *'utrolig'* (*incredible*) becomes *'utro'* (*adulterous*).

Improvements could be implemented to make SentidaV2 more directed towards opinion mining on social media. Here, an emoji-dictionary inspired by VADER could relatively easily be implemented. Also, a function that captures slang using multiple repetitions of the same letter – e.g. '*suuuuuper'* instead of '*super'*.

Furthermore, the values modulating the sentiment of sentences with '*men'* and '*dog'* (*but*), the values modulating the sentiment of sentences with exclamation marks, and the value for modulating the sentiment of words written in all capital letters are the same as the English SA-program VADER uses. They might not be generalizable to the Danish language and culture.

An expansion of the Emma validation set, both the number of sentences and the number of raters for each sentence, would increase the accuracy of the validation. An easy way of doing this would be to translate the English validation SST-2, as it is already rated and has been used before – this requires reflections on whether the sentiment scores are preserved through the translation, and whether this loss of accuracy is worth the saved resources. This would also enable more accurate comparison to the English SA tool benchmarks. The sentences would need ratings on the other three dimensions of the HSES model of Emma.

Besides containing 352 sentences rated for valence, Emma also contains the ratings of these sentences in the three other dimensions: Intensity, controllability, and utility. These four dimensions can be used to distinguish 16

discrete emotions (Trnka et al., 2016). With an expansion of Emma, the validation set can be used to create a tool for multidimensional sentiment analysis by using it as a training set for NNs that will be able to detect and distinguish these 16 emotions in written language.

The specific methods for training the neural networks are implementable with a basis in Google's BERT framework (Munikar et al., 2019) that is even more context-aware than SentidaV2 and might enable future studies to reliably recognize multidimensional aspect-based, context-aware, sentiment in Danish texts (Wang et al., 2016).

9 Conclusion

This paper introduces Emma (Emotional Multidimensional Analysis) and SentidaV2. SentidaV2 is the new state-of-the-art in Danish sentiment analysis and is shown to be significantly better than current methods ($p < 0.01$) in three different binary datasets with varying qualities of human coded sentiment scores. SentidaV2 correctly classified 82% of the sentences in a binary TrustPilot review dataset (TP2) and 70% of the sentences in a binary (positive and negative) dataset made from Emma. Emma is a completely new dataset for Danish sentiment analysis with 352 sentences scored in a four-dimensional circumplex emotional space by 30 coders using a citizen science approach with a custom application. Beyond the improvements to the validation process of sentiment programs in Danish, Emma also takes the field one step closer to multi-dimensional Danish emotional SA.

The study's main contribution is a novel multidimensional validation dataset for Danish SA that enables an array of future research possibilities regarding fine-tuning neural networks for multidimensional SA. The study also moves the Danish SA quality closer to international standards found in English and Chinese SA systems. There are some limitations in methodology regarding the size of Emma and the number of coders that warrants further research efforts. Future studies can focus on the utility of Emma in training neural networks for Danish multidimensional sentiment analysis and might enable Danish SA to exceed international standards in emotional classification of texts.

10 Acknowledgments

We would like to thank Fabio Trecca for the invaluable feedback and the team behind SENTIDA for helpful discussions and inspiration.

11 Public Access

SentidaV2 and the Emma validation dataset can be accessed through:

https://github.com/esbenkc/emma

12 References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606 [Cs]*. http://arxiv.org/abs/1607.04606

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699392084 11068

Fleiss, J. L., & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303 300309

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (Version

0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv:1402.3722 [Cs, Stat]*. http://arxiv.org/abs/1402.3722

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2017). *Learning Word Vectors for 157 Languages*. 5.

Heck, R., Vuculescu, O., Sørensen, J. J., Zoller, J., Andreasen, M. G., Bason, M. G., Ejlertsen, P., Elíasson, O., Haikka, P., Laustsen, J. S., Nielsen, L. L., Mao, A., Müller, R., Napolitano, M., Pedersen, M. K., Thorsen, A. R., Bergenholtz, C., Calarco, T., Montangero, S., & Sherson, J. F. (2018). Remote optimization of an ultracold atoms experiment by experts and citizen scientists. *Proceedings of the National Academy of Sciences*, *115*(48), E11231–E11237. https://doi.org/10.1073/pnas.1716869115

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. https://doi.org/10.18653/v1/P18-1031

Hutto, C. J. (2019). *Cjhutto/vaderSentiment* [Python]. https://github.com/cjhutto/vaderSentiment (Original work published 2014)

Hutto, C. J., & Gilbert, E. (2014, May 16). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media*. Eighth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *ArXiv:1607.01759 [Cs]*. http://arxiv.org/abs/1607.01759

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krippendorff, K. (2004). Content analysis: An introduction to its methodology Thousand Oaks. *Calif.: Sage*.

Lauridsen, G. A., Dalsgaard, J. A., & Svendsen, L. K. B. (2019). SENTIDA: A New Tool for Sentiment Analysis in Danish. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift*, *4*(1), 38–53.

Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Cambridge : Oelgeschlager, Gunn & Hain. http://archive.org/details/basicdimensionsf0000mehr

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Munikar, M., Shakya, S., & Shrestha, A. (2019). *Fine-grained Sentiment Classification using BERT*. 5.

Nielsen, F. Å. (2017, April 28). *AFINN*. AFINN. http://www2.compute.dtu.dk/pubdb/views/edoc_download.php/6975/pdf/imm6975.pdf

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois press.

Pedersen, M. K., Rasmussen, N. R., Sherson, J. F., & Basaiawmoit, R. V. (2017). *Leaderboard Effects on Player Performance in a Citizen Science Game*. 8.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv:1802.05365 [Cs]*. http://arxiv.org/abs/1802.05365

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, *89*(4), 344–350. JSTOR.

R Core Team. (2013). *R: A language and environment for statistical computing* [R]. R Foundation for Statistical Computing. http://www.R-project.org/

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Trnka, R., Lačev, A., Balcar, K., Kuška, M., & Tavel, P. (2016). Modeling Semantic Emotion Space Using a 3D Hypercube-Projection: An Innovative Analytical Approach for the Psychology of Emotions. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00522

Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 225–230. https://doi.org/10.18653/v1/P16-2037

Watson, D., & Tellegen, A. (1985). Toward a Consensual Structure of Mood. *Psychological Bulletin*, *98*(2), 219–235. https://doi.org/10.1037/0033-2909.98.2.219

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231–240. https://doi.org/10.1519/15184.1