Emma meets SENTIDA: Danish Computational Analysis of Emotion in Text

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

Esben Kran (201909190@uni.au.dk)

Søren Orm H. (201907685@post.au.dk)

University of Aarhus, Department of Cognitive Science

1 Abstract

In this paper, we report the current state of Danish sentiment analysis and the process of creating the new state-of-the-art Danish sentiment analysis (SA) tool by updating the current state-of-the-art SENTIDA to V2. SentidaV2 shows significant improvement in classifying sentiment in text compared to SENTIDA ($p < 0.01$) in three different validation datasets. We also present a new domain general validation dataset called Emma 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. It consists of many coders with a wide demographic representation and sentences scored in a four-dimensional emotional circumplex space 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, Emma

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 data. 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.

2.1 Research question

How can the current state-of-the-art in Danish sentiment analysis be improved?

From this question, two hypotheses are tested.

H1 Sentiment scoring accuracy can be increased through higher syntactical and semantic awareness in SENTIDA.

H2 By creating a new, coded validation dataset for SA, we can approach practical multi-dimensional Danish emotional SA.

2.2 Sentiment Analysis (SA)

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 differs 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; Labille et al., 2017; 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) uses a BoW approach with limited syntactic awareness with a unidimensional scale of valence based on the circumplex model of affect (Russell, 1980) from -5 to 5 that indicates the level of positive emotion associated with specific words.

These are generally only on one or two dimensions despite the circumplex models development 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.

2.3 Current Danish lexical SA

The first Danish SA AFINN was developed from an interest in Twitter sentiment analysis (Nielsen, 2011) and the Danish version is developed from machine translations of English sentiment lexicons checked by F. Nielsen (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 are separately rated by the three authors and the mean value between the three is used as the score (Lauridsen et al., 2019). Additionally, values not overlapping between AFINN and SENTIDA are pulled from AFINN and re-rated using the SENTIDA coding scheme.

2.4 Difficulties in sentiment analysis

2.4.1 Bag of Words (BoW) approach

Current tools use a semi-BoW approach to sentiment analysis, where a word is associated with the relevant sentiment score and added to the total sum of sentiment in the text. The BoW approach is limited in several ways:

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.

2.4.2 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 has a different sentiment 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.

2.5 Quantitative dimensional measures of emotion

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 totally consists of four dimensions for emotion: 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 (Mäntylä et al., 2018; Wang et al., 2016).

2.6 Current databases

Current databases extensively document valence in texts but there are very few in Danish. Notable examples are listed below.

- ANEW: Affective norms for English words with 1034 words rated on 9-point scales of valence and arousal (Bradley & Lang, 1999)
- Expansion of the ANEW dataset with demographic information, another dimension of emotional dominance, and extension to 13,915 English lemmas (Warriner et al., 2013)
- A new ANEW dataset consisting of words labelled specifically for use on Twitter (Nielsen, 2011)
- Aarup: The Danish dataset used in SENTIDA's scoring mechanism (Guscode, 2019/2019; Lauridsen et al., 2019)
- AFINN-X: Datasets in Danish, Swedish, French, and English used in AFINN's scoring mechanism (Nielsen, 2017, 2015/2019)
- The NRC Word-Emotion Association Lexicon is a compilation of 14,182 unigrams (words) rated on a binary sentiment of positive or negative and 4-class emotional associations to anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Mohammad & Turney, 2010, 2013).
- A Taiwanese emotional association lexicon from 2007 that drew down 5,000,000< blog posts and gave words an emotional association based on the emoticons used in the text (Yang et al., 2007).
- 4,300 Dutch words rated on valence, arousal, dominance, and age of acquisition (Moors et al., 2013)
- SST-2 and SST-5. Stanford Sentiment Tree with binary and pentamery sentiment scores of movie reviews (Devlin et al., 2019)

These represent an array of different methods for collecting this data. Beyond these lexicon-based sentiment databases, there is an array of pre-trained natural-language processing models (e.g. BERT (Devlin et al., 2019), FastText (Xu & Du, 2019), etc.) that constitute hidden vectors trained from giant databases on different languages. Some of these include sentiment analysis training based on data that resembles the validation dataset Emma introduced in this paper (Hoang et al., 2019; N. Liu et al., 2019; Maas et al., 2012; Munikar et al., 2019; Rana & Cheah, 2016).

2.7 Social gamification and cognitive incentive using citizen science

Emma's database is rated by volunteers using citizen science (CS). A gamified meta-system is implemented to engage the CS coders. The modelling of the data collection tool is not focused on gamification as different personality types receive motivation from different factors (Anus & Melle, 2014; Orji et al., 2014) and a bad fit leads to negative results. Emma uses a leaderboard-based system that increases amount of ratings per coder (Alsawaier, 2018; Mekler et al., 2017) despite some academic criticisms (Pedersen et al., 2017). Coder interviews confirms incentive.

2.8 Improvements introduced

Despite the drastic improvement to Danish SA implemented in AFINN and SENTIDA, they are still

limited compared to the English state-of-the-art BoW tool VADER (Hutto & Gilbert, 2014). Additionally, some aspects of syntactical awareness are missing from all these tools.

The current standard in Danish SA validation tests is missing high reliability and domain generality. TrustPilot scores associated with reviews are the current standard where large irregularities are present. A comprehensive list of tool and validation updates in this paper are listed below.

Python implementation of SENTIDA; data collection tool for Danish machine learning data annotation; validation set on four emotional dimensions for future neural network-based Danish multidimensional SA; *Ikke* (*not*) synonyms introduction and *ikke* in questions; *Men* (*but*) intensity updates; exclamation mark intensity modification; uppercase intensity modification; word context awareness area updates; binary tests on three validation datasets; and implementation of a rigorous test process with 1,000s of iterations.

3 Validation text data collection tool

The validation text data collection tool (VTDC) was developed using the Google Apps Script (GAS) to develop interaction between Google Sheets and Google Forms as a Danish substitute for data annotation for supervised machine learning like Amazon MTurk as it does not include demographic representation in Denmark (Paolacci & Chandler, 2014). Additionally, it's inexpensive compared to other programs (Djenno et al., 2015).

VTDC consists of three parts: The survey document, the database, and the GAS file. The database is online, Google Sheets-based, and includes three sheets: *Sentiment logging*, *leaderboard*, and *validation texts*. Respectively, these consist of all annotated sentences (allowing for duplicates) with demographic information for each, texts scored with ID, and the texts needing annotation in eight different analytically interesting categories.

The GAS is the interface unit between the database and the survey form that reacts to the event trigger from the form *onFormSubmit*. This enables it to activate a function every time someone submits the form. It acts as follows:

- Pulls an updated texts list from the database
- Generates a (hash)table of these sentences
- Reads and saves response with demographic information (ID, GDPR, public, region, age, education, occupation) into the *sentiment logging* sheet
- Updates gamification aspects in the form: ID points, leaderboards, and database sorting, excluding people who do not want their ID to be public
- Selects 20 random sentences and inputs them in the form

The Form 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 coders. The coding scheme is structured into the four different emotional dimensions as seen in Table 1 - Coding scheme for 4D emotional text annotation.

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 1 - Coding scheme for 4D emotional text annotation

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. See more information about the theory behind the structure in Social gamification and cognitive incentive using citizen science.

3.1 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, education levels, occupations, and ages as previous tools do not have a large representation of demographic variety with all coders being WEIRD and male (Lauridsen et al., 2019; Nielsen, 2017).

The demographic variety of the 30 coders spans nearly all levels of society ranging from jobless to self-employed, middle school to Ph.D., 17-65 years old, and from all official regions of Denmark.

3.1.1 Intercoder agreement

Measurement of the intercoder reliability is performed using a oneway pairwise interclass correlation coefficient (ICC) test average (Fleiss & Cohen, 1973; Weir, 2005). The test uses a oneway algorithm that assumes row effects random intercepts for each rater which means the raters are assumed to vary in their own subjective standards for text ratings. The test uses the intercoder agreement in ICC. Agreement offers a more strict but precise output compared to consistency. Based on this measure, the ICC of the raters in Emma is 0.743. The quality of a coefficient of 0.743 is defined as *good* by Cicchetti, 1994, but as *moderate* on modern scales that focus on better reproducibility (Koo & Li, 2016).

In Sentida's dataset, the interrater reliability was measured using Krippendorff's alpha at a value of 0.667 (Lauridsen et al., 2019) which is defined as the smallest acceptable value with 0.8< defined as a good measure (Krippendorff, 2004).

Based on the data collection method and the compromises made to reach mass appeal with the dataset, the ICC is relatively strong for Emma compared to previous datasets though improvement is still possible.

3.2 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 are equals the mean values of ratings for that sentence given by the coders.

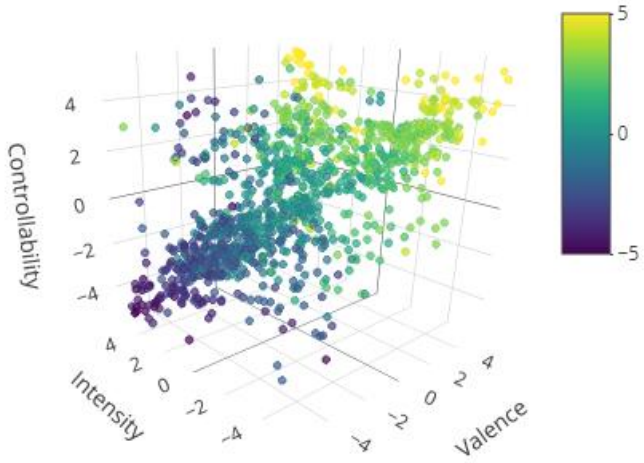

Figure 1 - Sentences and their ratings in the four dimensions (color: utility)

4 SentidaV2 – the Python Tool

The SentidaV2 python script takes a sentence and splits it into individual words but 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. The Danish

words/slang: *Ikke*, *ik*, *ikk*, *intet*, *aldrig*, and *ingen* meaning *not*, *never*, and *nothing* are common negators used in Danish. If any of these words are identified in a sentence, the sentiment of the word before and the following three words are multiplied by –1. However, if a question mark at the end of a sentence is detected, the negation is removed.

Another feature of SentidaV2 is the *men*-detection and updating functionality. If the Danish word *men* (*but*) is identified in a sentence the sentiment in the part of the sentence before *men* is multiplied by 0.5 and the part after by 1.5. These values are the same used by the English sentiment analysis program VADER (Hutto, 2014/2019).

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. The values are the same used in VADER (Hutto, 2014/2019).

If a word is written in all capital letters, the sentiment of that word is multiplied by 1.733 based on empirical tests (Hutto, 2014/2019). If a word in the sentence passed into SentidaV2 is in the list of intensifiers, the sentiment of the word before and the following three words are multiplied by the magnitude corresponding to the intensifier word in the list.

5 Validation datasets

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

6 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 2). 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.

7 Predicting Sentiment

Below, in Table 2, 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 $p_{Emma} < 2.2e\text{-}16$).

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

8 Is SentidaV2 better?

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 any means of splitting them from each other. 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, we see 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.

9 Is Emma better?

None of the SA programs perform as well on Emma as they do on the two Trust Pilot validation sets. A few reasons might be the cause for this difference in performance. First, the Trust Pilot 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 sentence 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.

10 International state-of-the-art

Comparing results of SA-programs across languages poses a challenge as the datasets are different or need to be translated. The best parallel between Danish and English SA benchmarks is TP2 and SST-2 as they both consist of reviews and are highly used in the respective fields. Figure 2 shows the current state-of-the-art in English SA outlined. The best model is the $BERT_{LARGE}$ that reaches a 93.1% binary accuracy on the SST-2 dataset. Compared to SentidaV2's accuracy of 82%, there is still a big difference. This difference can be reduced in future research utilizing Emma for training neural networks such as BERT like Munikar et al., 2019.

Model	SST-2	
	All	Root
Avg word vectors [9]	85.1	80.1
RNN [8]	86.1	82.4
RNTN [9]	87.6	85.4
Paragraph vectors [2]	–	87.8
LSTM [10]	–	84.9
BiLSTM [10]	–	87.5
CNN [11]	–	87.2
$BERT_{BASE}$	94.0	91.2
$BERT_{LARGE}$	**94.7**	**93.1**

Figure 2 - Current international state-of-the-art on SST (Stanford Sentiment Tree) binary movie review validation set (from Munikar et al., 2019)

11 Future research

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

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; 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 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. These sentences would then need ratings on the other three dimensions of the model.

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).

12 Conclusion

This paper introduces Emma 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 and confirms the first hypothesis (H1). SentidaV2 correctly classified 82% in a binary TrustPilot review dataset (TP2) and 70% in a sentiment binary coded sentence 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 for supervised machine learning data collection. The second hypothesis (H2) is partly confirmed as Emma, beyond the improvements to the validation process, also takes us 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 pre-training 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.

13 Acknowledgments

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

14 Public Access

SentidaV2 and the Emma validation dataset can be accessed through:

https://github.com/esbenkc/emma

15 References

Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. *The International Journal of Information and Learning Technology*, *35*(1), 56–79. https://doi.org/10.1108/IJILT-02-2017-0009

Anus, S., & Melle, I. (2014). Diagnosis and differentiating instruction: Differentiated task assignment in chemistry education. *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research For Evidence-Based Teaching and Coherence in Learning. Part*, *3*, 219–228.

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

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report No. C-1). *Gainesville, FL: NIMH Center for Research in Psychophysiology, University of Florida*.

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

Djenno, M., Insua, G. M., & Pho, A. (2015). From paper to pixels: Using Google Forms for collaboration and assessment. *Library Hi Tech News*, *32*(4), 9–13. https://doi.org/10.1108/LHTN-12-2014-0105

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

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/10.1177/001316447303300309

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.

Guscode. (2019). *Guscode/Sentida* [R]. https://github.com/Guscode/Sentida (Original work published 2019)

Hoang, M., Bihorac, O. A., & Rouces, J. (2019, September 30). Aspect-Based Sentiment Analysis using BERT. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. https://www.aclweb.org/anthology/W19-6120.pdf

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*.

Labille, K., Gauch, S., & Alfarhood, S. (2017). *Creating Domain-Specific Sentiment Lexicons via Text*

Mining. 8.

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.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, N., Shen, B., Zhang, Z., Zhang, Z., & Mi, K. (2019). Attention-based Sentiment Reasoner for aspect-based sentiment analysis. *Human-Centric Computing and Information Sciences*, *9*(1), 35. https://doi.org/10.1186/s13673-019-0196-3

Maas, A. L., Ng, A. Y., & Potts, C. (2012). *Multi-Dimensional Sentiment Analysis with Learned Representations*.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

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

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525–534. https://doi.org/10.1016/j.chb.2015.08.048

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

Mohammad, S. M., & Turney, P. D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. 9.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *ArXiv:1308.6297 [Cs]*. http://arxiv.org/abs/1308.6297

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169–177. https://doi.org/10.3758/s13428-012-0243-8

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

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv:1103.2903 [Cs]*. http://arxiv.org/abs/1103.2903

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

Nielsen, F. Å. (2019). *Danish resources*. https://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6956/pdf/imm6956.pdf

Nielsen, F. Å. (2019). *Fnielsen/afinn* [Jupyter Notebook]. https://github.com/fnielsen/afinn (Original work published 2015)

Orji, R., Vassileva, J., & Mandryk, R. L. (2014). Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Modeling and User-Adapted Interaction*, *24*(5), 453–498. https://doi.org/10.1007/s11257-014-9149-8

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

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

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.

Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: Comparative analysis and survey. *Artificial Intelligence Review*, *46*(4), 459–483. https://doi.org/10.1007/s10462-016-9472-z

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

Shafie, A. S., Sharef, N. M., Azmi Murad, M. A., & Azman, A. (2018). Aspect Extraction Performance with POS Tag Pattern of Dependency Relation in Aspect-based Sentiment Analysis. *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 1–6. https://doi.org/10.1109/INFRKM.2018.8464692

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

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

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

Xu, J., & Du, Q. (2019). A Deep Investigation into fastText. *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 1714–1719. https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00234

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Building Emotion Lexicon from Weblog Corpora. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 133–136. https://www.aclweb.org/anthology/P07-2034