Vol 5, No. 1, 2020

Journal of language works

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, contact@esbenkc.com
Søren Orm, 201907685@post.au.dk, sorenorm@live.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

Abstract

Sentiment analysis (SA) is the research and development field of computationally analysing emotion in text. One usage example of SA could be to track the sentiment of a company's mentions on Twitter or to analyse a book's positivity level. In this paper, we attempt to add to this work in two ways. First, we further develop the current tool Sentida (Lauridsen et al., 2019), which was originally developed to score valence in text. Valence is the amount of positivity in a text, e.g. a review. Our new version has a higher awareness of punctuation and syntax compared to the earlier version and shows significant improvement in classifying valence compared to the previous version in three different validation datasets ($p < 0.01$). Second, we develop a test dataset which future developers of SA can use called Emma (Emotional Multidimensional Analysis). In Emma, we supplement the dimension valence with a further three emotional dimensions: Intensity, dominance, and utility in a dataset of sentences scored by human coders on these four dimensions. The emotional dimensions are based on cognitive psychology work throughout the last 65 years.

With Emma, we present both a more reliable validation dataset and the possibility of further improving the Danish SA field by using the dataset to train a neural network with machine learning for analysing more complex emotions in text. The current standard is the 1-dimensional classification of positivity in text, but with this approach, we allow for a classification in the four dimensions of the Emma dataset that reveals much more complex emotions in texts. To allow others to work with Sentida and Emma, we help update the currently available Sentida optimized for Python and publish Emma on Github.

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

1. Introduction

To discover new patterns and 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. Computers have proven remarkably efficient in processing data and discovering patterns. One problem, however, is that 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 sentiment analysis tools that can assess the positivity of texts for Danish. These tools can be improved to better match international standards and expanded so that they are able to assess the complex range of emotions that people express in written language. This paper seeks to further develop the field of Danish computational linguistics.

How can the current state-of-the-art in Danish sentiment analysis (SA) be improved? This paper attempts to answer this question in two ways. First, we further develop the current tool, Sentida (Lauridsen et al., 2019), which was originally developed to score *valence*. Valence is defined as the amount of positive sentiment in a text. Sentida currently stands as the state-of-the-art of Danish SA. The improvements introduced in this paper focuses on incorporating higher syntactical and semantic awareness to increase the accuracy, i.e. the degree to which the score produced by the SA tool corresponds to the interpretation by human readers. Secondly, we introduce a dataset, Emma (Emotional Multidimensional Analysis), consisting of sentences scored by human coders on four different emotional dimensions. These dimensions are *valence* (positivity), *intensity*, *dominance*, and *utility*. They are based on cognitive psychology work on expressions of emotions conducted throughout the last 65 years (e.g. Hepach et al., 2011; Osgood et al., 1957; Russell, 1980; Trnka et al., 2016).

2. Sentiment Analysis

SA is a part of applied computational linguistics and attempts to quantify the emotions, most often positivity, of written language, especially on the internet or in large corpora of texts like newspaper databases. Examples of use cases are extracting the positivity of political articles to analyse specific newspapers' political leanings (Enevoldsen & Hansen, 2017) or to understand customers' feelings regarding companies in reviews on TrustPilot, e.g. '*dårlig oplevelse'* (*bad experience*) giving a sentiment score of -0.33 or *'jeg er meget tilfreds'* (*I am very satisfied*) giving a valence score of 0.72. In the field of SA, sentiment is the presence of negative or positive charge in a word, sentence, or larger piece of text negative or positive charge in (B. Liu, 2012; Mäntylä et al., 2018). Approaches to analysing sentiment differ widely in complexity. From bag-of-words approaches (BoW) where the sentiment of the input is determined by matching words to a sentiment lexicon, to aspect-aware neural network (NN)-based approaches achieving advanced context awareness influencing the same word's sentiment score based on its context (Hoang et al., 2019; N. Liu et al., 2019) and 2-dimensional valence-intensity (VI) SA with combinations of NN techniques (Maas et al., 2012; Wang et al., 2016). Below we expand on the differences between the three.

2.1 Bag of Words (BoW) approach

Many current SA tools use a semi-BoW approach to sentiment analysis, where a word is associated with a sentiment score (**Fejl! Henvisningskilde ikke fundet.**) (Hutto & Gilbert, 2014), irrespective of the context of the word. The aggregate sentiment scores of the words in the text are then used as an indication of how positive the text is. This approach is computationally efficient but has some limitations as outlined below. The state-of-the-art in English BoW SA is the VADER (Hutto, 2014/2019; Hutto & Gilbert, 2014) while the Danish SA field has the tools AFINN (Nielsen, 2011, 2017, 2015/2019) and Sentida (Guscode, 2019/2019; Lauridsen et al., 2019). They work in roughly the same way.

Table 1 - Example of lexicon words with sentiment score

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

There are four main problems of only using BoW for SA in text. These mainly arise from a missing context awareness. First, it ignores the syntactical relationship between words in the text. Relationships like verb-noun structure are ignored and generalized, which limits the accuracy as seen in Table 2, where *'animal'* and *'you'* defines how *'wild'* is interpreted. Secondly, it ignores adverbs and negation words like *'extremely'* and *'not'* (see Table 2). Thirdly, it does not reflect human sentiment perception. Humans use pattern recognition and context knowledge to understand the emotions of a text compared to just looking up in a dictionary (Hasson et al., 2020) which leads to different interpretations of the sentiment for the same word. And fourth, there is no difference in the rating of homographs as sentiments are only matched to strings of letters, as seen in the fourth example in Table 2, where *'lyst'* means two different things but is equal in textual representation.

Table 2 - Examples of each problem with BoW

Problem	*Example (Danish)*	*Example (English)*
Syntactical awareness	Du er for *vild*! Det er et *vildt* dyr.	You are *wild*! It's a *wild* animal.
Intensity modification	Du er *smuk*. Du er *ekstremt smuk*. Du er *ikke smuk*.	You are *beautiful*. You are *extremely beautiful*. You are *not beautiful*.
Context awareness	Han er *vild*. Bogen er *vild*.	He is *crazy*. The book is *crazy*.
Homographs	Jeg har *lyst* til *lyst* kød	I *want light* meat

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 that add contextual understanding. In this paper, Sentida's methods are also improved to minimize these problems.

AFINN is a pure BoW method that also includes emoticons while Sentida is an expansion on the AFINN word list and includes simple negations and simple punctuation awareness with a multiplier if the sentence includes an exclamation mark. VADER is the most advanced and used English BoW tool today and includes an array of context awareness functionality, e.g. the amount of exclamation marks decides multiplication, fully upper case words increase the intensity, emoticons are included, and the sentiment is modulated based on *'but'* in the sentence. These will be described in more depth along with examples in our description of the updated Sentida introduced in this paper.

2.2 Neural network approaches

Many sophisticated modern approaches to SA use a neural networks (NN) approach. 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 some of the state-of-the-art NNs for sentiment analysis.

The advantage of using neural networks is that they automatically find patterns in the text in the same way humans do (Hasson et al., 2020). Some neural networks have for example implemented aspect extraction to perform so-called aspect-based sentiment analysis (ABSA) (Hoang et al., 2019; N. Liu et al., 2019; Rana & Cheah, 2016; Shafie et al., 2018). In ABSA every single word is assigned different sentiment scores based on the context they appear in. This can solve problems such as the lack of syntactical awareness and the homograph problem described above (Table 2). The way neural networks are trained (which we will not go into here) also allows them to perform limited inference of the context of text which gives them better awareness when performing SA.

The only problem of the BoW approach that can be fixed without abandoning the approach altogether is the intensity modification while the other problems are easier to solve with neural networks. Even though NN-based architectures have high accuracy, their processing speed is very slow, and a lot of good quality data (approximately 25.000 sentences) is required to train them. This creates problems with the workload of the dataset development as well as the time for processing the text itself during text analysis. The reason for using BoW in practice is that they are generally faster.

2.3 Multidimensional SA

Additionally, SA analysing both the valence (positivity) *and* the intensity (described below) of the sentiment in text have been developed using NN. The current state-of-the-art uses combinations of convolutional NN and long short-term memory algorithms (Wang et al., 2016) with training datasets coded for valence and intensity. With a large enough dataset, the same can be achieved on the four dimensions of Emma to perform emotional analysis in Danish texts which is the purpose of developing the Emma dataset. In this paper, the first steps towards a larger dataset with the four dimensions of *valence*, *intensity*, *controllability*, and *utility* are taken. By training a neural network with these sentences, it will be possible to score a text on each of these parameters. An example from the dataset is the sentence *'Jeg endte tit med at sidde inde på kontoret og tude.'* (*I often ended*

up sitting in the office crying) which has a valence score of -0.87, an intensity score of 0.73, a controllability score of -0.67, and a utility score of -0.73 with scores between -1 and 1. With an NN-based architecture, this would then be generalizable to new sentences that are not in the dataset. Below is a description of dimensional emotion classification.

2.4 Potential of quantitative dimensional SA with Emma

Beyond Ekman's basic emotions of anger, disgust, fear, happiness, sadness, and surprise from traditional psychology of emotional expression (Ekman, 1992), the focus in Emma is on the dimensional quantitative models of emotion PAD (Mehrabian, 1980) and the hypercubic (four dimensional) semantic emotion space (HSES) (Trnka et al., 2016).

The model used for Emma consists of four dimensions of emotions: 1. Valence, 2. Intensity, 3. Controllability, 4. Utility (Mehrabian, 1980; Trnka et al., 2016). They are the result of improvements on earlier models (Osgood et al., 1957) and introduce new capabilities to modern SA in the previously described multi-dimensional sentiment analysis (Poria et al., 2018; Wang et al., 2016). PAD describes emotion in the scales of pleasure, arousal, and dominance and the HSES model describes emotion in the scales of valence, intensity, controllability, and utility.

Valence represents the positive associations of the target word or text, e.g. *'sur*' (*angry*) having a negative valence score and *'glad'* (*happy*) having a positive valence score. *Intensity* is how intensely the emotion is represented, e.g. *'okay'* (*okay*) compared to *'fantastisk'* (*fantastic*). *Controllability* is how in-control one feels in the represented emotion, e.g. '*frustration*' is uncontrollable while '*happiness*' is controllable. *Utility* is how beneficial or harmful the emotion is, e.g. '*happiness*' being beneficial and '*sadness*' being harmful.

The potential of the multidimensional models in SA is the ability to identify different emotions in text beyond the normal one-dimensional positivity or valence scale as described earlier. The HSES model uses four dimensions. It is thereby able to define 16 discrete emotions with different combinations of values on the four scales (Trnka et al., 2016) with a fine-grained accuracy. Some emotions are not discernible with less than these four dimensions. By having all four dimensions of the HSES model in Emma, it is possible to create models that are 2D, 3D, and 4D to perform fine-grained analysis of the sentiment in text.

When improving Sentida, this paper only looks at the valence dimension of the Emma dataset for validation, but Emma's other dimensions and the insights from the dataset itself are valuable for future research in the described neural network SA research field.

2.5 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). Researchers have been interested in Twitter SA because it allows us to get an everyday view of positivity regarding specific topics extracted through searches on Twitter. An

example is to search for #dkpol on Twitter, which is the hashtag many Danish politicians use, to get an overview of the current political sentiment.

Sentida is a lexicon consisting of the 5,263 most-used Danish sentiment-carrying lemmas (Lauridsen et al., 2019). These words were separately rated on a valence scale of -5 to 5 by the three authors, and the mean rating was used as the lexicon valence 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. Additionally, the stems of words were used to extend the lexicon's range to approximately 35,000 Danish words in total.

AFINN and Sentida (Lauridsen et al., 2019; Nielsen, 2017) both use the previously described BoW approach with limited syntactic awareness. Furthermore they operate with only one scale of emotion, viz. valence which is based on the circumplex (coordinate system-based) model of affect (Russell, 1980). Beyond the valence dimension, the circumplex model of affect also concerns itself with the *intensity* of the emotion (Russell, 1980). Additionally, as described above, there are other circumplex 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 *valence* and *arousal* scales (Trnka et al., 2016). These are representative of frameworks with a more nuanced description of emotion. The next step in emotional SA should incorporate these. As mentioned, we have implemented this coding in the Emma corpus, and we will return to the method applied in coding Emma below. When it comes to testing our improvements on Sentida, however, we will simplify the presentation by discussing only *valence*.

3. Improvement process

3.1 Sentida

The updated Sentida tool, like AFINN and Sentida takes a sentence, splits it into individual words and saves the order of the words. It matches the individual words in the sentence with a list of valence-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 word. As described earlier, Sentida is not context-aware beyond the one sentence and does not understand the real-world context of the text which limits it compared to humans. In this paper, we improve on Sentida by adding several intensity modifiers such as *'ikke'* (*not*), including synonyms and abbreviations, e.g. *'ik'* and *'ikk'* (*not*), *'aldrig'* (*never*), and *'ingen'* (*none*). In the example below, the sentence would get a score of -2.3 despite having the word *'godt'* (*good*) in it because of the word *'aldrig'* (*never*). In the previous model, *'aldrig'* (*never*) would not negate the sentiment.

"Det er **aldrig** (-1 x →) godt (+2.3)." ⇒ sentiment score: -2.3

"That is **never** (-1 x →) good (+2.3)." ⇒ sentiment score: -2.3

If the synonyms of *'ikke'* (*not*) appear in questions, there is no negation – in Danish, the usage of *not* in a question does not negate the sentiment. The original Sentida negates for all *'ikke'* (*not*) no matter the context.

> "Er det ikke (~~-1 x →~~) godt (2.3)**?**" ⇒ sentiment score: +2.3
>
> "Isn't (~~-1 x →~~) that good (2.3)**?**" ⇒ sentiment score: +2.3

We often see *'but'* in a sentence changing the intensity of the words preceding the *'but'* compared to the words proceeding it. If there is *'men'* (*but*) in a sentence, the part of the sentence after *'but'* carries more sentimental charge than the part before *'but'*. The English SA 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). These values are also used in the updated Sentida.

> "Maden (+0.3) var god (+2.3), (← x 0.5) **men** (1.5 x →) serviceringen (+0.3) var elendig (-4.3)." ⇒ 1.3 - 6 ⇒ sentiment score: -4.7
>
> "The food (+0.3) was good (+2.3), (← x 0.5) **but** (1.5 x →) the service (+0.3) was horrendous (-4.3)." ⇒ 1.3 - 6 ⇒ sentiment score: -4.7

In text, especially informal, exclamation marks (EM) are often used as intensifiers. For each 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)" ⇒ sentiment score: 4.6
>
> "It is so cool (+3.6)**!** (← x 1.291)" ⇒ sentiment score: 4.6

Capital letters can have a similar function to exclamation marks by increasing the sentiment of words. If a word is written in all capital letters, the sentiment of that word is multiplied by 1.733. This value is the same used in VADER (Hutto, 2014/2019):

> "DET ER SÅ SEJT (+3.6). (← x 1.733)" ⇒ sentiment score: 6.2
>
> "IT IS SO COOL (+3.6). (← x 1.733)" ⇒ sentiment score: 6.2

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 natural-language processing field (NLP, the field concerned with computational language analysis) and more performant. Writing Sentida in Python thus eases the process of incorporating improvements made for other languages and using Sentida with other NLP tools.

3.2 Emma

Beyond being useful for future research in multidimensional SA, Emma is also introduced as a new validation dataset using the valence scale of the scored sentences. 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 Danish SA programs mainly because it is easy to acquire a large set of rated sentences.

However, using TrustPilot reviews has its problems: Spelling mistakes often occur and the SA program will not be able to recognize the words; rating mistakes happen, causing the validation to lose accuracy, like writing *"god service"* (*good service*) and giving a rating of one star; and there is no clear etiquette for how to write or rate reviews, e.g. a 4-star review might describe why the product got 4-stars or 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.

Contrary to TrustPilot reviews, Emma is based on ratings by 30 raters, who all received the same instructions on how to rate the sentence. With a specifically coded dataset, Emma is more controlled than TrustPilot. It is also designed to be more naturalistic in nature than reviews and has a representative syntactical sample. In total, 352 sentences were rated. The sentences belong to categories such as **simple negative sentences**, **complex positive sentences**, **sentences with *'men'* (*but*),** and **sentences with negations**. To ensure a naturalistic representation in the sentences, they were selected from Danish news articles and Danish commentary on social media. Examples include *"den globale økonomi ryster ikke"* (*the global economy is not shaking*) and *"jeg er i tvivl, om de besøgende turister er begejstret"* (*I am not sure if the visiting tourists are excited*), both complex sentences that represent a variety of sentiment context information.

The raters were recruited using citizen science (CS) to ensure a broad demographic representation of Danes. CS also contributes to motivation for the coders as it attempts to motivate through their sense of assisting in a scientific endeavour which has been shown to increase engagement (Heck et al., 2018; Pedersen et al., 2017). The coders were given a description of the four dimensions similar to the presentation in 3.3 along with some of the same examples.

3.3 Emma data collection program

The program is composed of a form where the coders rate sentences, after which these ratings are automatically sent to a database and updated with new sentences for the next coder (see the form here: forms.gle/MkFp28pCyphGFHMP7). The form includes the coding scheme and is designed to

be easy to use and understand so that it does not seem intimidating for the coders. The coding scheme is structured into the four different emotional dimensions as seen in **Fejl! Henvisningskilde ikke fundet.**.

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

Coders are then asked to rate 20 sentences on each dimension from -5 to 5 with neutral = 0. These are shown after an introduction to CS, the project itself, and the mechanics of the survey. Furthermore, a top five leader board is displayed as a gamification incentive. Despite some scientific debate (Alsawaier, 2018; Mekler et al., 2017; Pedersen et al., 2017), leader boards have often been proven to incentivize coders to give more ratings. After finishing their coding, the respondents are asked to provide information about themselves such as a username, region, age, educational level, and occupation. If the username has been used before, we assume that it is the same person and add the points for this session to the previous session to reflect their position on the leader board.

3.4 Coders

To ensure as wide of a demographic representation as possible 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), which might lead to a skew in the emotions associated to different sentences caused by the differences in upbringing and cultural expectations.

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

3.5 Intercoder agreement

An interesting aspect of the ratings is the important metric of intercoder agreement (IA). This signifies how much the different CS coders agreed with each other in their ratings. An IA approaching 0 signifies that the coders do not agree with each other, while one closer to 1 signifies that they *do* agree with each other on the emotional association for the sentences. To calculate the

IA, we use Krippendorff's alpha (Krippendorff, 2004) like the Sentida team (Lauridsen et al., 2019). Coders with too few overlaps with the other coders were excluded from the test.

The IA is represented by Krippendorff's alpha value for each dimension and was calculated using the IRR package (Gamer et al., 2019) in R (R Core Team, 2013). The alpha for valence in Emma is 0.4, 0.15 for intensity, 0.28 for controllability, and 0.25 for utility. For comparison, the Sentida dataset's three coders obtained a Krippendorff's alpha value of 0.667 (Lauridsen et al., 2019) on over 5000 words, which is an *acceptable* value, with an alpha value above 0.8 defined as a *good* measure by Krippendorff (Krippendorff, 2004).

We see a much lower IA in Emma than in the Sentida dataset. This might be caused by the size of Emma, the naturalistic ambiguity of the sentences' emotional correlation, and the different perspectives of the coders caused by the demographic variety.

3.6 Ratings

The ratings returned from Emma defines each sentence as a dot in a four-dimensional space representing the hypercubic definition of emotional space (Trnka et al., 2016) (**Fejl! Henvisningskilde ikke fundet.**). The coordinates for each sentence correspond to the mean values of ratings for that sentence given by the coders. The three geometric axes represent controllability, intensity, and valence, while the colour represents utility. If a sentence e.g. has a valence of 2, an intensity of -4, a controllability of 3.5, and a utility of 2, the sentence would be represented as a dot with the coordinates (2, -4, 3.5) in the coordinate system and have a light green colour.

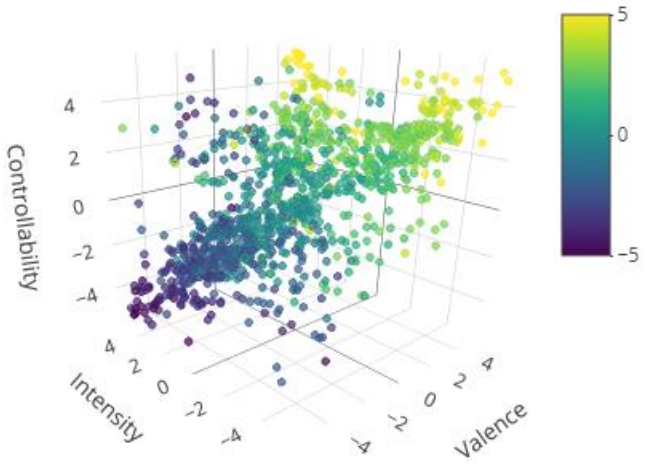

Figure 1 - Sentences and their ratings in the four dimensions (colour: utility)

Looking at the graph, we see that there might be correlations in the different dimensions. In Table 3, the correlation between the different scores is plotted. We see that every single emotional dimension is correlated with every other emotional dimension. This is in contrast with the HSES paper that notes a limited amount of colinearity between the four dimensions. This is the reason that

Trnka et al. propose that it is valid to include the different dimensions describing emotion (Trnka et al., 2016).

There is a high colinearity between different parameters, in particular between *valence*, *controllability* and *utility*. This indicates that the coders perceive controllability and utility to be more-or-less synonymous with valence, and it is in effect an undermining of the HSES model's two last parameters for this study. When there is such a collinearity, the first two parameters can be used to describe the other two which limits the utility of having them. Despite Emma's intercoder agreement being low, this still calls for further studies as the uniqueness of each dimension has been documented (Trnka et al., 2016) but seems to be missing statistical validity in our data. It undermines the use of the four-dimensional model but as we only use the valence dimension that describes several of the others, it is not a problem in this study.

Table 3 - Pairwise correlation table: Spearman Correlations

Dimensions			Spearman's rho	p
Valence	-	Intensity	-0.447	< .001
Valence	-	Controllability	0.719	< .001
Valence	-	Utility	0.892	< .001
Intensity	-	Controllability	-0.395	< .001
Intensity	-	Utility	-0.440	< .001
Controllability	-	Utility	0.751	< .001

In the following, we will test the improvements to Sentida presented in 4.1 above against the Emma corpus. We test only on the more robust dimension of *valence*.

4. Test results

The precision of our updated Sentida was tested using three validation sets: A corpus of TrustPilot reviews previously used to validate Sentida (Lauridsen et al., 2019) with only lowercase and no punctuation (TP), a different set of TrustPilot reviews with casing and punctuation (TP2), and the valence dimension of the Emma corpus introduced above. TP and TP2 consist of respectively 7019 and 7015 reviews from the website TrustPilot along with the number of stars 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. TP is included to compare with the Sentida paper, and TP2 is included to ensure that all the improvements made in Sentida, including the scoring of exclamation points and the use of upper case, are utilized.

4.1 Validation process

To assess how good the updated Sentida is at classifying sentiment in sentences compared to other Danish SA-programs, we first labelled the sentences in TP, TP2, and Emma. For TP and TP2, the sentences got the label positive, if the review had received 4 or 5 stars, and negative, if the review

had received 1 or 2 stars. The middle fifth of Emma's 11 rating levels (-1.1 to 1.1) for valence were removed in the same way and the sentences were labelled positive if the sentiment ratings were above 1.1, and negative if the ratings were below -1.1. We then processed the reviews in TP, TP2, and Emma with Sentida, with and without the updates, and AFINN to assign sentiment score.

The sentiment scores from TP, TP2, and Emma for each SA program were each first split up into two parts. The first part, containing 75% of the sentiment scores, was used to make a model (called a logistic regression) that guessed whether the sentence should be classified as negative or positive. This model can be thought of as a way to define a limit. If the sentiment score of a sentence is above this limit, the model will guess that the sentiment of the sentence is positive. If the sentiment is below the limit, the model will guess that the sentiment of the sentence is negative.

The remaining 25% of the sentences were used to test how good the models were. This was done by letting the model guess if the person had given the company a positive or negative rating on TrustPilot, or by letting it guess if the Emma sentences were scored negatively or positively by the coders. It then calculated the percentage of correct guesses. We split the data to test the model's accuracy on data it has not 'seen' in the 75% dataset.

Splitting the dataset like this may create selection bias and widely differing accuracies which is why the datasets were split and the accuracies calculated 1,000 times. The average accuracy and the 95% confidence interval were extracted for each dataset (Table 4). A t-test was then used on the accuracies to determine whether there was any difference between Sentida with and without the updates.

4.2 Predicting Sentiment

Fejl! Henvisningskilde ikke fundet. summarises the average accuracies and the average 95% confidence intervals of the three SA-programs, AFINN, Sentida, and the updated Sentida over the 1000 tests, on the three validation sets. On average, the updated Sentida was found to have a significantly higher accuracy at binarily classifying whether the sentiment of the sentences in TP was positive or negative (M = 0.8063, SD = 0.007), in TP2 (M = 0.8183, SD = 0.008), and in Emma (M = 0.69577159, SD = 0.0420486), 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 = 0.67486016, SD = 0.04352). The t-values are $t_{TP}(1997.2) = 3.2837$, $t_{TP2}(1988.6) = 98.488$, and $t_{Emma}(1988.897.3) = 10.99550.813$, and the differences are significant for all datasets ($p_{TP} = 0.001$, $p_{TP2} < 2.2e\text{-}16$, and $p_{Emma} < 2.2e\text{-}16$).

Table 4 - Results from accuracy tests (TP: TrustPilot, TP2: TrustPilot 2, EM: Emma)

	TP	TP: 95% CI	TP2	TP2: 95% CI	EM	EM: 95% CI
Updated Sentida	0.8063	0.7891 to 0.8226	0.8183	0.7999 to 0.8365	0.7159	0.5922 to 0.8194
Sentida	0.8052	0.788 to 0.8215	0.7817	0.7623 to 0.8015	0.6016	0.4743 to 0.7915
AFINN	0.7497	0.731 to	0.7494	0.7284 to	0.5022	0.3771 to

	0.7676	0.7695	0.6271

In Table 4, the columns TP, TP2, and Emma (EM) display the average of the different SA programs' accuracies calculated as described above. The updated Sentida can for example be seen to guess right on TP2 nearly 82%, compared to the original Sentida's 78 %. (For comparison, chance would be 50% correct guesses). The columns TP: 95% CI, TP2: 95% CI, and EM: 95% CI display the 95% Confidence Intervals (CI) for the different datasets. The 95% confidence interval means that we are 95% sure that the true average lies within this range. This uncertainty arises from the fact that the different ways of splitting the datasets give us different accuracies.

5. Evaluation of Emma and the updated Sentida

The updated Sentida has the smallest advantage on the TP dataset. This was as expected. All the words are lower case, meaning that the update to include capitalization in Sentida had no effect. There is no punctuation in the reviews, meaning no punctuation effects like *'?'* and *'!!'* could be used. 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 a sentence-to-sentence-basis. We see a statistically significant difference between the performance of Sentida with and without updates, but the difference is still 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 in Sentida can be tested, i.e. exclamation points and capital letters. Presumably, therefore the difference in performance is especially prominent when compared to the performance of Sentida without the new improvements on TP2.

The same pattern is observed for the Emma sentences; a quite substantial and significant difference was found between the accuracy of Sentida before and after the improvements. 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. The reason for this difference may be that the sentences in Emma display a more complex and context dependent usage of language, not necessarily having an obvious positive or negative sentiment. In the TrustPilot reviews the context is given, i.e. people write about their experiences with a product often explicitly positively or negatively. Emma can be said to better reflect real-world contexts. The updated Sentida is less successful at extracting sentiment from the complex texts, i.e. it scores higher on the simpler TP datasets. The difference in accuracy is less striking between TP2 and Emma for the updated Sentida than it is for the original Sentida and for AFINN.

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 and give better credibility to the intercoder agreement numbers.

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.

6. Future research

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

For example, improvements could be implemented to make Sentida more directed towards opinion mining on social media. Here, an emoji-dictionary inspired by VADER could relatively easily be implemented, and a function that captures slang using multiple repetitions of the same letter – e.g. *'suuuuper'* 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 or across different genres, and it would therefore make sense to test these values with Danish sentences and Danish populations instead of using the English basis to ensure cultural validity.

In addition, Sentida 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 35,000 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, e.g. *happy* and *happiest*, and some words become grossly mis-rated, e.g. the word *'utrolig'* (*incredible*) becomes *'utro'* (*adulterous*).

As mentioned, 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 dataset SST-5, 215,154 phrases rated by 3 coders (Socher et al., 2013). There is a potential loss in accuracy that should be weighed against the saved resources on if the phrases should be recoded for Danish. It would grant a better comparison to the English standards that use this as a benchmark. The sentences would need further ratings on the other three dimensions of Emma, i.e. intensity, controllability, and utility depending on the usefulness of these dimensions.

To raise the intercoder agreement of the dataset, there are several routes to take. First, the dataset can be expanded to get a more accurate representation of intercoder agreement. Second, the sentences can be selected to be easier to code. Third, the coders' demographic can be homogenized to decrease sociocultural variation between understanding of words' sentiment. Fourth, the survey's

framing of each dimension and instructions can be made clearer. The second and third solutions to increase intercoder agreement might reduce how naturalistic the dataset is which should be discussed for future research.

Beyond containing 352 sentences rated for valence, Emma also contains the ratings of these sentences in three other emotional dimensions: Intensity, controllability, and utility. These four dimensions have been shown to enable distinguishing 16 discrete emotions (Trnka et al., 2016) but our data shows that this might not be valid because of collinearity between the emotional dimensions. This warrants further research and creates doubt in the use of more than two dimensions for emotional detection in text, according to our collinearity analysis of the coders' ratings of the dimensions.

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 possibly distinguish these 16 emotions in written language as described previously. Briefly touching on the possible development methods, the basis can be implemented in Google's more context-aware BERT framework (Munikar et al., 2019) with the added use of the Danish Gigaword project (Strømberg-Derczynski et al., 2020) and might enable future studies to reliably recognize multidimensional aspect-based, context-aware sentiment in Danish texts if the practical value of multidimensional sentiment is studied further.

7. Conclusion

This paper introduces Emma (Emotional Multidimensional Analysis) and introduces improvements to the current state-of-the-art Danish sentiment analysis (SA) tool, Sentida. Emma is a completely new dataset for Danish sentiment analysis with 352 sentences scored in four dimensions: Valence, intensity, controllability, and utility. These dimensions were chosen based on previous work in cognitive psychology and allows coding to distinguish between 16 different emotions using these four dimensions. The sentences are rated by 30 coders using a citizen science approach with a purpose-built data collection program. By introducing a dataset that is not based on TrustPilot reviews, Emma also takes the research field one step closer to multi-dimensional Danish emotional SA of a wider range of texts. However, further research on the value of the multi-dimensional approach is needed.

With the new improvements, Sentida is significantly improved compared to its previous version ($p < 0.001$) in three different datasets with varying qualities of human coded positivity scores for texts and sentences. For the sentences in a TrustPilot review dataset (TP2), Sentida could correctly guess whether a review from the dataset was positive or negative in 82% of the cases. Sentida was also able to correctly guess whether the sentences in the dataset Emma had received a positive or a negative scoring 71% of the times. In particular, the latter is a sizable improvement over previous SA tools, as the previous version of Sentida had an accuracy of 60% on the Emma dataset.

The study's main contribution is a novel multidimensional 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, the multi-dimensional approach, and the number of coders that warrants further research efforts. Future studies can focus on the utility of the multi-dimensional approach and expansion of Emma for training neural networks in Danish multidimensional sentiment analysis and might enable Danish SA to exceed international standards in *emotional* classification of texts.

Acknowledgments

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

Public Access

Sentida is available on https://github.com/guscode/sentida.

The Emma dataset is available on https://github.com/esbenkc/emma.

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

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*.

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/10.1080/02699939208411068

Enevoldsen, K. C., & Hansen, L. (2017). *Analysing Political Biases in Danish Newspapers Using Sentiment Analysis*. 12.

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

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

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

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

Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing Emotions Along the Dimensions of Valence, Arousal, and Communicative Frequency – Implications for Social-Cognitive Tests and Training Tools. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00266

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

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.

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

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)

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/

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

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. 12.

Strømberg-Derczynski, L., Baglini, R., Christiansen, M. H., Ciosici, M. R., Dalsgaard, J. A., Fusaroli, R., Henrichsen, P. J., Hvingelby, R., Kirkedal, A., Kjeldsen, A. S., Ladefoged, C., Nielsen, F. Å., Petersen, M. L., Rystrøm, J. H., & Varab, D. (2020). The Danish Gigaword Project. *ArXiv:2005.03521 [Cs]*. http://arxiv.org/abs/2005.03521

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