|  |
| --- |
| 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](mailto:201909190@post.au.dk), [contact@esbenkc.com](mailto:contact@esbenkc.com))  Søren Orm ([201907685@post.au.dk](mailto:sorenorm@live.dk), [sorenorm@live.dk](mailto: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. A 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, by collecting sentences scored by human coders on four different emotional dimensions in a dataset called Emma (Emotional Multidimensional Analysis). These dimensions are valence, intensity, dominance, and utility and are based on cognitive psychology work throughout the last 65 years. Second, we further develop the current tool Sentida which was originally developed to score the first dimension of Emma, valence, in text. Valence is the amount of positivity in a text. Our new version has a higher awareness of punctuation and syntactics 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).

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 through 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 each dimension of the Emma dataset that allows us to see 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

# Introduction

In order 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. A 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 tools that can assess the positivity of texts for Danish. These tools can be improved, not only to better match international standards; they can also be expanded so 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 in this aspect.

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

# Sentiment Analysis

SA is a part of applied computational linguistics and attempts to quantify the emotions, most often positivity and negativity, 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*) or *'jeg er meget tilfreds'* (*I am very satisfied*) giving a valence score of -0.33 and 0.72 respectively.In the field of SA, sentiment is the emotion present in a word, sentence, or larger piece of text and valence is the negative or positive charge (B. Liu, 2012; Mäntylä et al., 2018). Approaches to analysing sentiment differ widely in complexity from a bag-of-words approach (BoW), where the sentiment of the input is determined by matching words to a sentiment lexicon; to aspect-aware neural network (NN)-based approaches with advanced context awareness influencing the same word’s sentiment score based on its context (Hoang et al., 2019; N. Liu et al., 2019); and 2-dimensional valence-intensity (VI) SA using combinations of NN techniques (Maas et al., 2012; Wang et al., 2016). Below we expand on the differences between the three.

## 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 (Table 1) (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 tool 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 that 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 (Table 2). Secondly, it ignores adverbs and negation words (Table 2). Thirdly, it does not reflect human sentiment perception. We 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). And fourth, there’s no difference in the rating of homographs as it is only matched to the lexicon (Table 2).

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. In this paper, Sentida’s methods are also improved to limit the effect of these problems.

## Neural network approaches

Many sophisticated modern approaches to SA use an 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), where a single word is assigned different sentiment scores based on the context which can solve problems such as 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 a better meta context awareness when performing SA.

When talking of the problems of the BoW approach, what can be fixed within the scope of still using BoW is only the intensity modification while the other problems are easier to solve with neural networks. The reason for then still using BoW instead of neural networks is that it’s generally faster. 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 which creates problems with the workload of the dataset development as well as the time for processing the text itself during text analysis.

## Multidimensionel 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 with 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.

This 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 with Emma. By training a neural network with these sentences, it will be able to tell you what score the text receives 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*) that is scored with 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 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.

## Potential of quantitative dimensional SA with Emma

Beyond Ekman’s basic emotions of anger, disgust, fear, happiness, sadness, and surprise from traditional psychology (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 introduces 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 or negative associations of the target word or text, e.g. *‘sur* (*angry*) having a negative valence score and *‘happy’* (*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 often is controllable. Utility is how beneficial or harmful the emotion is, e.g. with happiness being beneficial and sadness often being harmful.

The potential of these models in SA is the ability to identify different emotions in text beyond the normal one-dimensional positivity or valence scale as described earlier. The HSES model validated using four dimensions because they were able to define 16 discrete emotions with different combinations of values on the four scales (Trnka et al., 2016) with a fine accuracy. Without the four dimensions, some of the emotions were not able to be identified. By having all four dimensions of the HSES model in Emma, it is then possible to create models that are 2D, 3D, and 4D to perform fine-grained analysis of 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 interesting for future research in the described neural network SA space.

## 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). The reason for being interested in Twitter SA is that 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 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.

The current standard in Danish (Lauridsen et al., 2019; Nielsen, 2017) builds on the previously described bag of words (BoW) approach with limited syntactic awareness, and with a scale of valence based on the circumplex (coordinate system-based) model of affect (Russell, 1980). This valence scale varies from -5 to 5 and indicates the level of negative vs. positive emotion associated with specific words.Beyond the valence dimension, the circumplex model of affect also concerns itself with the intensity of the emotion (Russell, 1980) where the current tools nearly only concern themselves with the one dimension of valence. Additionally, as described above, 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) are representative of frameworks with a more nuanced description of emotion. The next step in emotional SA should incorporate these.

# Improvement process

## 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, it is not context-aware beyond the one sentence and doesn’t understand the real-world context of the text which limits it compared to humans.In this paper, we improve on the available tool, expanding Sentida by adding several intensity modifiers such as *‘ikke’* (*not*) synonyms and abbreviations, e.g. *‘ik’* and *‘ikk’* (*not*), *‘aldrig’* (*never*), and *‘ingen’* (*none*). In this example, the sentence would get a score of negative 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 →~~) forkert (-2.6)**?**” ⇒ sentiment score: -2.6

“Isn’t (~~-1 x →~~) that wrong (-2.6)**?**” ⇒ sentiment score: -2.6

We often see *‘but’* in a sentence that changes 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, concerned with computational language analysis) field 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.

## 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, which is a problem because some might in a 4-star review write why the product got 4-stars, while others might write why it didn’t get 5-stars. The same experience with a product might cause two different consumers to write similar reviews while rating the product differently, too.

As opposed to the TrustPilot reviews, Emma 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 its 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 or complex positive sentences along with syntactical sentences such as sentences with *‘men’* (*but*) and sentences with negations. The raters were recruited using citizen science to ensure a broad demographic representation of Danes. Citizen science 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).

## 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. The form can be seen on this link (Kran & Orm, 2020). The form includes the coding scheme and is designed to be easy to use and understand so it does not seem intimidating for the citizen science (CS) coders. The coding scheme is structured into the four different emotional dimensions as seen in Table 3.

|  |
| --- |
|  |

Table 3 - Coding scheme for the citizen science raters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 the scheme, Citizen Science, the project itself, the mechanics of the survey, and a top five leader board for the a gamification incentive functionality is displayed. By using a leaderboard, we provide further incentive for the CS coders which is proven to incentivize the coders to give more ratings despite some scientific debate (Alsawaier, 2018; Mekler et al., 2017; Pedersen et al., 2017). After the coding process, the respondents are asked to provide information about their demographic such as a username, region, age, educational level, and occupation. If the username has been used before, we assume that it’s the same person and add the points for this session to the previous session reflect their position on the leader board.

## 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 student to employers) and job titles, educational backgrounds (middle school to Ph.D.), age ranges (17-65 years), and location (all five Danish regions are represented).

## Intercoder agreement

An interesting aspect of the ratings is the important metric of intercoder agreement or reliability. This signifies how much the different CS coders agreed with each other in their ratings. An intercoder agreement towards 0 signifies that the coders with different social, geographical, political, and age backgrounds do not agree with each other while one closer to 1 signifies that they *do* agree with each other on the emotional association to the sentences. A limitation to the metric calculated below is that the coders with too few overlaps with the other coders on which sentences they coded were not included in the intercoder reliability test.

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

The intercoder agreement in Emma is therefore not very large but respectable enough for us to be able to use the mean values of the four aspects of each sentence as optimal codings. It also signifies that there is not a large discrepancy in how different groups in Denmark view the emotional aspects of sentences.

## Ratings

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

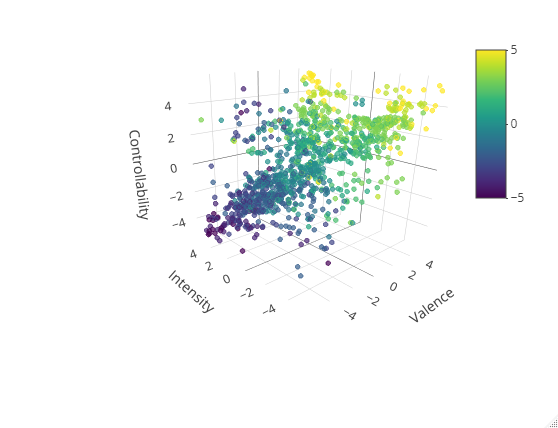


Figure 1 - Sentences and their ratings in the four dimensions (color: utility)

Looking at the graph, we see that there might be correlations in the different dimensions. In Table 4, 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 which is the reason that the different dimensions are valid to include in describing emotion (Trnka et al., 2016).

There is a high colinearity between different parameters. This indicates an undermining of the HSES model’s two last parameters by a high correlation between the last two dimensions (controllability and utility) and the first two (valence and intensity). When there is such a colinearity, the first two parameters can be used to describe the other two which limits the utility of having them. However, it is still interesting to look at because the 16 emotions are uniquely describable in the HSES model using these dimensions (Trnka et al., 2016).

Other interesting aspects arise from the correlations themselves. For example, high positive emotion indicates that people rate it as low intensity, high controllability, and a reliably high utility. In the same vein, high intensity then indicates low controllability and utility.

Table 4 - Pairwise correlation table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Spearman Correlations** | | | | | | | | |
|  |  |  |  |  |  | **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 |
|  |  |  |  |  |  |  |  |  |
| \* p < .05, \*\* p < .01, \*\*\* p < .001 | | | | | | | | |

# 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 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. 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.

## 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 convert the reviews and sentences from text to a 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 letting it guess if the Emma sentences were scored negatively or positively by the coders. Then calculating 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 might 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 5). A t-test was then used on the accuracies to determine whether there was any difference between Sentida with and without the updates.

## Predicting Sentiment

In Table 3 the average accuracies and the average 95% confidence intervals of the three SA-programs, AFINN, SentidaSENTIDA, and the updated SentidaV2 over the 1000 tests, on the three validation sets are summarised. 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 tTP(1997.2) = 3.2837, tTP2(1988.6) = 98.488, and tEmma(1988.897.3) = 10.99550.813, and the differences are significant for all datasets (pTP = 0.001, pTP2 < 2.2e-16, and pEmma < 2.2e-16).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5 - 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 0.7676 | 0.7494 | 0.7284 to 0.7695 | 0.5022 | | 0.3771 to 0.6271 |

In Table 5, the colomns TP, TP2, and Emma (EM) display the average accuracies for the different SA programs’ accuracy calculated as described above. The updated Sentida can for example be seen to guess right on TP2 nearly 82% of the time while chance would be 50%. 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 gives us different accuracies.

# Evaluation of Emma and the updated Sentida

Using the validation set TP poses a few problems. All the words are lower case, meaning no effect from capitalization. There is no punctuation in the reviews, meaning no punctuation effects like *‘?’* and *‘!!’*. Additionally, each review consists of multiple sentences without punctuation which makes it hard to split them up. This is a problem in sentences containing ‘*men’* (*but*) because the sentiment modulation is only meant to be applied on a sentence to sentence basis. We see a 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, this is why the difference in performance is especially prominent when compared to the performance of Sentida without the new improvenemts on TP2 as was the intention.

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 might 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 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. Emma 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.

# 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 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 5263to 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 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, e.g. cultural differences, differences in homographs and homonyms between the two languages, and the quality of the translation could along with many other factors lower the accuracy of the dataset. This potential loss of accuracy has to be weighed against the saved resources of not having to rate sentences. Implementing a translated version of SST-2 would also enable more accurate comparison to the English SA tool benchmarks. In addition to the before-mentioned difficulties with translating the sentences, the sentences would also need ratings on the other three dimensions of Emma, i.e. intensity, controllability, and utility.

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 as described previously. Briefly touching on the possible development methods, the basis can be in Google’s BERT framework (Munikar et al., 2019) that is even more context aware than Sentida and might enable future studies to reliably recognize multidimensional aspect-based, context-aware sentiment in Danish texts (Wang et al., 2016).

# Conclusion

This paper introduces Emma (Emotional Multidimensional Analysis) and 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 that showed that it is possible to distinguish between at least 16 different emotions using these four dimensions. The sentences are rated by 30 coders using a citizen science approach with a novel data collection program. Beyond the improvements to the validation process of sentiment programs in Danish, by introducing a dataset that is not based on TrustPilot reviews (see DISC), Emma also takes the field one step closer to multi-dimensional Danish emotional SA.

With the new improvements, Sentida is significantly improved compared to its previous version (p < 0.01) 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.

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 and the number of coders that warrants further research efforts. Future studies can focus on the utility 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.

# Author bios

Søren Orm is affiliated with the Aarhus Cognitive Science program and is interested in cognitive neuroscience and human behaviour.

Esben Kran is affiliated with the Aarhus Cognitive Science program and is interested in utilizing machine learning and neuroscience technologies.

# Acknowledgments

We would like to thank Fabio Trecca for the 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*.

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

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

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

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

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

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

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

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

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

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

Kran, E., & Orm, S. (2020, January 4). *Forskningsprojekt: Følelser i tekst*. Google Docs. https://docs.google.com/forms/d/e/1FAIpQLSeyHaD3nFVTu0nQe3pLMbCXwrc3tBS4vpSPUWPBXIDOn54gCQ/viewform?usp=embed\_facebook

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.

Poria, S., Hussain, A., & Cambria, E. (2018). *Multimodal sentiment analysis /*. Springer.

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

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

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

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

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