Vol. 4, No. 1, 2019

Journal of language works

SENTIDA: A New Tool for Sentiment Analysis in Danish

Gustav Aarup Lauridsen, 201804481@cc.au.dk
Jacob Aarup Dalsgaard, 201809346@post.au.dk
Lars Kjartan Bacher Svendsen, 201807375@post.au.dk
University of Aarhus

Abstract

In the midst of the Era of Big Data, tools for analysing and processing unstructured data are needed more than ever. Being among these, sentiment analysis has experienced both a substantial proliferation in popularity and major developmental progress. However, the development of sentiment analysis tools in Danish has not experienced the same rapid development as e.g. English tools. Few Danish tools exist, and often the ones available are either ineffective or outdated. Moreover, authoritative validation tests in low-resource languages, are missing, which is why little can be deduced about the competence of current Danish models. We present SENTIDA, a simple and effective model for general sentiment analysis in Danish, and compare its competence to the current benchmark within the field of Danish sentiment analysis, AFINN. Combining a lexical approach with several incorporated functions, we construct SENTIDA and categorise it as a domain-independent sentiment analysis tool focusing on polarity strength. Subsequently, we run different validation tests, including a binary classification test of Trustpilot reviews and a correlation test based on manually rated texts from different domains. The results show that SENTIDA excels across all tests, predicting reviews with an accuracy above 80% in all trials and providing significant correlations with manually annotated texts.

Keywords: Sentiment analysis, Danish, automated text analysis, Data Linguistics, SENTIDA

"A word after a word after a word is power."
Margaret Atwood

1. Introduction

Until recently, one would usually buy a travel magazine or ask a friend to give his or her recommendations before planning a trip to Paris. With the Digital Revolution and the introduction of the World Wide Web, things have changed substantially with numerous blogs, online fora and social network services now providing individuals with the opportunity to gain insights and advice from strangers. Additionally, there is a rapid increase in the number of people making their opinions available to others via the Internet, making it a trend that is probable to continue skyrocketing in the future (Pang & Lee, 2005).

Journal of language works

According to Surowiecki (2004), the collective opinions of a community can be more informative than the opinions of a few experts. This means that the countless amounts of opinions on the Web will often provide you with a more confident idea about trends, political movements – and what you should do in Paris – than experts or magazines.

Sentiment analysis enables effective and precise interpretations of collective opinions, facilitating automated text data analysis in areas spanning from research to business. With one click, sentiment analysis can analyse thousands of Tripadvisor reviews and help decide whether you should spend your afternoon in the City of Love admiring the view from The Eiffel Tower or enjoying the fine arts at The Louvre.

1.1 Sentiment analysis

Sentiment analysis, also known as 'opinion mining' is a tool within the field of NLP, i.e. Natural Language Processing, that is used for identifying the tone conveyed by a text. The approach of sentiment analysis is based on the assumption that words contain sentiments, and thus, it operates by looking at the words of a given text to then calculate an overall sentiment score for the text. That is, one uses sentiment analysis to detect and categorise the emotional content expressed in a given text by analysing the individual words, which is why it is generally defined as the "computational treatment of opinion, sentiment and subjectivity in text" (Pang & Lee, 2008:1). For the purpose of this article, a 'text' is any coherent arrangement of words, and it can range from a single word to an entire library. To infer the sentiment of a text, a sentiment lexicon consisting of words and their respective sentiment ratings is needed. This so-called lexical approach operates by looking up the words in a text one by one to retrieve sentiment properties which give either an accumulated score of the words in the text or a mean score per word. The lexical approach belongs to a general approach called bag-of-words which describes approaches that do not consider the order or syntax of words in a sentence. The limitations of the bag-of-words technique is discussed further in Section 2.3.

In a scientific context, sentiment analysis has, among other things, been used to predict the onset of depression (Choudhury, Gamon, Counts, & Horvitz 2013) and developmental changes on the stock market (Bollen, Mao, Zeng, 2011). Furthermore, it is widely used in business, where companies collect increasing amounts of data about their customers' sentiments (Liu, 2012).

1.2 Existing tools in Danish

The AFINN lexicon is the most substantial contribution to the field of Danish sentiment analysis so far (Nielsen, 2018). Constructed by associate professor from DTU, Finn Årup Nielsen, the lexicon consists of 3552 words which have been manually rated on a discrete interval valence scale ranging from -5 (negative) to +5 (positive). Even though it is mainly aimed towards sentiment analysis of microblogging (e.g. tweets), AFINN has also shown to be useful within other fields. For example, a study undertaken by Enevoldsen & Hansen (2017) used AFINN to detect political bias in Danish newspapers.

Journal of language works

Nonetheless, a thorough examination of the wordlist of the AFINN lexicon revealed several significant limitations. First of all, many of the words were discovered to be inflections (e.g. it includes six inflections of the word *fattig* 'poor'). This diminishes the total number of unique words markedly and leaves AFINN with a smaller vocabulary of 3089 words. Moreover, in a random sample of 200 words, three of the words were found to be highly misrated, e.g. *udsigtsløs* 'hopeless'/'futile' has a score of +2.

Finally, another limitation to AFINN is that all the words have been rated solely by Nielsen, which means that the ratings lack robustness and are prone to a certain level of subjective idiosyncrasies (Nielsen, 2018).

Another matter of concern with AFINN is that it is yet to be validated, making it hard to deduce anything particular about its competence. This missing validation means that a frame of reference is non-existent, why insecurities are attendant when using the tool.

1.3 Motivation

The above sections emphasise the importance of developing a new efficient sentiment analysis tool in Danish that would help analysts, scientists, politicians and companies benefit from the vast amounts of Big Data. The observed gap between the potential of Big Data and the analysis tools available is the driving force of this paper, in which we will propose a new solution for Danish sentiment analysis.

We will, more specifically, present a computational sentiment analysis tool that consists of 1) a sentiment lexicon with a focus on both the polarity of words and the strength of the sentiment expressed by these; 2) a function to stem words, which artificially expands our lexicon; and 3) functions to identify and take grammatical properties such as modifiers and negations into account.

In other words, we have devised a new tool for sentiment analysis in Danish which we have named *SENTIDA*.

2. The construction of SENTIDA

2.1 The SENTIDA lexicon

SENTIDA is a lexicon comprised of the existing Danish sentiment lexicon AFINN and a list of new words. The list of new words comes from DSL (Det Danske Sprog- og Litteraturselskab), and contains the 10.000 most frequently used lemmas in Danish, i.e. the canonical form of words (DSL lemmas, 2018). This resource has been created using the 2016 corpus *BAKSPEJLET*, comprising text material from different domains in the period from 1983 until 2016.

As SENTIDA is intended to be made domain-independent, this resource was found to be relevant because it asserts that selected words are indeed the words that most accurately characterise the written Danish language irrespective of domain. From the initial list of 10.000 lemmas, a subset was created containing only the lemmas with the following part-of-word tags; adjectives, adverbs,

Journal of language works

common nouns, verbs and interjections, as these are the ones that typically carry sentiment (e.g. the adjective *happy* carries a sentiment, whereas the preposition *about* does not). The new list containing 9.779 lemmas was then manually annotated independently by each of the three authors. As the lexicon aims to be comparable to the existing Danish polarity lexicon AFINN, it was decided that the way of scoring the lemmas had to be similar to that of Nielsen (2011). This means that SENTIDA is a polarity strength lexicon similar to AFINN.

The next step was creating a coding scheme for rating the lemmas. The proposed coding scheme consisted of an 11-point scale like AFINN for which a score of 0 meant neutral, a score of -5 meant very strong negative emotion and a score of 5 meant very strong positive emotion (see figure 1 below).

Very strong negative emotion	Strong negative emotion	Slightly negative emotion	Weak negative emotion	Very weak negative emotion	Neutral emotion	Very weak positive emotion	Weak positive emotion	Slightly positive emotion	Strong positive emotion	Very strong positive emotion
-5	-4	-3	-2	-1	0	1	2	3	4	5

Figure 1. The SENTIDA coding scheme

The lemmas were rated by the three authors, whereafter an inter-coder reliability score was calculated to ensure confidence in making replicable and valid inferences about the performance of the lexicon. The inter-coder reliability between the independent ratings was computed using Krippendorff's alpha, implemented in R Studio and the package *irr* (Gamer, 2015).

The resulting Krippendorff's alpha of 0.667 means that it is acceptable to make tentative conclusions using the lexicon (Krippendorff, 2004). The strong alpha score indicates that the scores of the three raters are congruent but not entirely unison.

After the mean sentiment score from the three raters was calculated for each lemma, all lemmas with an accumulated sentiment score of 0 were omitted from the lexicon.

Once the initial lexicon had been composed, it was compared to the list of words in the AFINN lexicon. The lemmas present in AFINN but missing in SENTIDA, amounting to a total of 1379, were then gathered, inspected and re-annotated using the same coding scheme as before. Once all the lemmas had been scored, the lexicon was assembled.

2.2 WordStem

To further expand the sensitivity of the SENTIDA lexicon, all of the lemmas were stemmed, using the Snowball stemmer provided in the R package *SnowballC* (Bouchet-Valat, 2015). What is meant by stem, in this case, is transforming the lemma into the part of the lemma that never changes but still carries the meaning of the original lemma. This way a lemma like *spille* '(to) play' would be stemmed into *spil* 'play', which would then be able to convey the meaning for any word with this stem such as *spiller* 'player', *spillende* 'playing', *spillerne* 'the players' and *spillet* 'the game'.

Journal of language works

This was followed by a subsequent aggregation of identical stems. That is, lemmas having the same stem were given the mean score of these identical stems. In most cases, the aggregation would include stems from all inflections of a specific word. However, in few instances words with different semantics have the same stem (e.g. *adskille* '(to) separate' and *adskillige* 'several'). This is a limitation to stemming in general.

As a result, this new compressed lexicon is now more sensitive to sentiment in texts because it can match the stem in the lexicon with any word with the same stem. More specifically, the compression creates an artificial expansion of the tool, as the target texts are also stemmed, allowing the tool to detect numerous inflections of a given word in a text. The lexicon was compressed to 5263 entries, detecting ~35.000 Danish words in total.

As evident in Figure 2, the lexicon has a slight bias towards stems with a positive sentiment (2839, corresponding to 54%) compared to negatively sentimented stems (2424, corresponding to 46%). Additionally, the majority of the stems are centered around 0, with approximately 53% of the stems carrying a sentiment score between -1 and 1. Conversely, it is observed that the lexicon contains more highly negative stems than highly positive stems, with 11% of the stems having a score between -3 and -5, whereas positive scores between 3 and 5 only account for 5% of the stems.

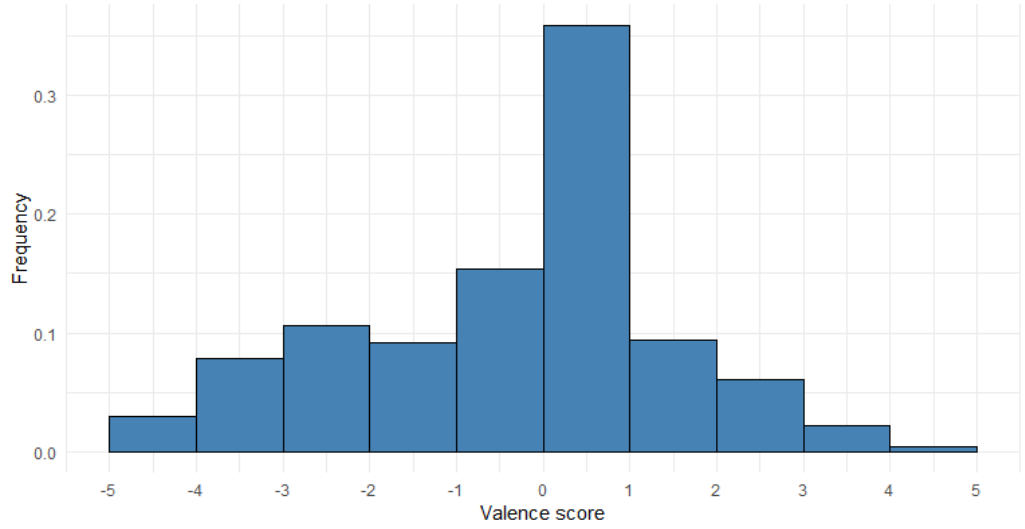

Figure 2: SENTIDA word score distribution

2.3 Three simple heuristics

Following the manual annotation process, we wanted to expand our sentiment analysis tool beyond the simple bag-of-words technique. In an attempt at simulating some degree of context-awareness, three simple heuristics were constructed: 1) negations reversing the polarity of perceived sentiment; 2) exclamation marks increasing perceived sentiment of sentences, regardless of polarity; 3) modifiers changing the perceived sentiment value of the following word in the text.

Journal of language works

2.3.1 Negations

One of the most dominant flaws in the bag-of-words approach is the failure to deal with negations. The following example from the Trustpilot validation corpus shows how common sentiment lexicons fail to deal with negations:

> "Det er ikke behageligt(+3)"
> "It's not comfortable(+3)"

As elaborated by Penka & Zeijlstra (2010), negations like the Danish word *ikke* 'not' tends to reverse the score of the upcoming sentiment-laden word. However, similar to English *not*, *ikke* is not only used as a negation in the Danish language, it is sometimes also used in conversation to prime a positive response to a question or to elicit a response to a statement. When used in questions, *ikke* occurs before the sentiment-laden word, whereas when used as a response-elicitor *ikke* occurs after the sentiment-laden word, e.g.:

> "Er det ikke vidunderligt herude?"
> "Is it not wonderful out here?"

and

> "Det er vidunderligt herude, ikke?"
> "It's wonderful out here, no?"

However, according to (Den Danske Ordbog, 2018), *ikke* is used far more often as a negation, and the use of *ikke* in a question primarily occurs in spoken language.

When using SENTIDA, the function will detect the word *ikke* and automatically reverse the scores of the two following words. This means that when *ikke* is used in a question before a sentiment-laden word, SENTIDA will wrongly reverse the score.

However, despite flaws, the function lifts SENTIDA beyond the bag-of-words principle substantially.

To illustrate how the function works we reuse the first example brought up in this section:

> "Det er **ikke** (-1 ×→) behageligt (+3)"
> "It's **not** (-1 ×→) comfortable (+3)"

This example shows how the reverser (-1) of the word, *ikke*, affects the score of the next word, *behageligt* (3), to create a final score of -1*3=-3.

2.3.2 Exclamation

When dealing with sentiment, some of the most modern tools have started to include the influence of exclamation marks. Penka & Zeijlstra (2010) researched the effect of exclamation marks on perceived sentiment. Through 30 reviews of 1.000 tweets, they found that an exclamation mark increased the perceived sentiment by an average of 29.10%. This effect has been incorporated in the SENTIDA function, modifying results as shown in this example from the Trustpilot test corpus:

Journal of language works

"Hvor uprofessionelt(-4)!(←×1.291)"
trustpilot.dk, uniquelady.dk

"How unprofessional(-4)!(←×1.291)"
trustpilot.dk, uniquelady.dk

In this example, the sentiment score of the text (-4) is multiplied by the exclamation mark intensifier (1.291), creating a final score of -5.164.

2.3.3 Adverb modifiers

The third heuristic is a function to deal with modifiers. Modifiers are words such as *extremely*, *incredibly*, or *mega*. These words serve the purpose of modifying the sentiment of an upcoming sentiment-laden word. Dragut and Fellbaum (2014) investigated the impact of adverbs on sentiment analysis, providing intensity modifying scores for several different modifiers.

Table 1: Effect of adverbs on sentiment ratings

Adverbs	Polarity change
meget (very)	1.2
mega (mega)	1.4
lidt (slightly)	0.8
ekstremt (extremely)	1.2
totalt (totally)	1.2
utroligt (incredibly)	1.2
rimelig (fairly)	1.2
seriøst (seriously)	1.4

In SENTIDA, some of these intensity modifiers have been translated and incorporated, while appropriate adverbs not present in Dragut and Fellbaum's (2014) list have been added as well. This translation is based on the assumption that the adverbs are used similarly in the two languages. Moreover, the translated intensity modifiers carry the same polarity change score as Dragut and Fellbaum's, whereas those added have been rated accordingly. The entire list of modifiers and their intensity modifier can be seen in Table 1.

Below is an example from the Trustpilot test corpus including an adverb modifier:

"Det er meget (1.2 ×→) irriterende (-2.667), at varen bliver leveret (1.58) for sent (-1)"
Trustpilot, coolshop.dk

"It's very (1.2 ×→) annoying (-2.667) that the package was delivered (1.58) too late (-1)"
Trustpilot, coolshop.dk

In this case, the word "irriterende"(*annoying*), with the score of -2.667, is multiplied by the polarity change score (1.2) of the intensity modifier "meget"(*very*) to create a final score of -3.2.

Journal of language works

2.4 The use of SENTIDA

Technically speaking, SENTIDA takes in a string of text, divides the text into words, and stems the words using the SnowballC stemmer (Bouchet-Valat, 2015). Consistent with the bag-of-words technique, each word is then matched against the SENTIDA sentiment lexicon, and if the word is present, the corresponding rating is added to the sentence score. If an *ikke* 'not' appears, the next two word scores are reversed. If one of the words is in the SENTIDA modifier list (see Table 1), the following word is multiplied by the corresponding multiplier. Similarly, if the text includes an exclamation mark, the sentence score is multiplied by 1.291.

The output of SENTIDA is the sentence score divided by the number of words contributing to the score, i.e. the mean score. Using the mean, as opposed to the accumulated score, is done in order to normalise the score with regards to the length of the text.

3. Validation

In a low-resource language like Danish, lacking in software tools and online resources for computational treatment of language, no corpora are available for testing sentiment analysis tools. In order to validate SENTIDA, a test corpus consisting of sentiment rated texts is needed in order to compare the sentiment scores of SENTIDA to those of human annotators.

Thus, three different test corpora were constructed simultaneously with the creation of SENTIDA; two corpora consisting of Trustpilot reviews and one corpus consisting of texts from different domains - that is, political texts, fictional literature and social media posts.

In the statistical testing of the tool, we use AFINN as a benchmark in order to provide a better validation for the use of SENTIDA.

3.1 Trustpilot validation corpora

Existing literature considers user-generated reviews to be an efficient source of sentiment rated texts. Being easily accessible and of great quantity, reviews are widely used when testing sentiment analysis tools, in that they can be used as a measure of how well a tool can predict human sentiment (Hutto & Gilbert, 2014). For the analysis of SENTIDA, 8800 reviews with an average of 74.6 words (SD = 78.9) were collected across categories from *trustpilot.dk*.

The corpus was then reorganised into two novel test corpora:

1. The first corpus includes 1-, 2-, 4-, and 5-star reviews, which are categorized into positive and negative reviews (1 & 2 = negative, 4 & 5 = positive). This corpus adds up to 7019 reviews.
2. The second corpus includes 1, 3, and 5-star reviews. This corpus adds up to 5307 reviews.

3.1.1 Logistic regression

Firstly, SENTIDA was tested against AFINN on the ability to predict the polarity of reviews in Corpus I using logistic regression as predictive analysis. This type of binary classification task is a standard way of testing sentiment analysis tools (Hutto & Gilbert, 2014). We interpreted the number

Journal of language works

of stars of the reviews as a binary dependent variable (1 & 2 = negative, 4 & 5 = positive) and then used SENTIDA's and AFINN's sentiment scores of the review text as predictor variables to create a logistic regression model for each lexicon. Both the SENTIDA model and the AFINN model were trained on 80% of the first Trustpilot corpus (consisting of 5615 reviews) and then subsequently tested on the remaining 20% (consisting of 1404 reviews). The underlying models used to test the accuracy of the lexicons both were as follows:

polarity of review ~ mean sentiment score of review

SENTIDA significantly predicted that a change in the mean sentiment score of a review had an effect in predicting the polarity of the review: $\chi 2(1) = 2422.8$, $p < .001$***

A positive change in the mean sentiment score had a significant positive impact on predicting a positive review: b = 2.49 (SE = 0.07), z = 35.57, $p < .001$***, odds ratio = 12.09

Similarly, AFINN significantly predicted that a change in the mean sentiment score of a review had an effect in predicting the polarity of the review: $\chi 2(1) = 1877.1$, $p < .001$***

A positive change in the mean sentiment score had a significant positive impact on predicting a positive review: b = 0.99 (SE = 0.03), z = 35.59, $p < .001$***, odds ratio = 2.69

The results suggest that both AFINN and SENTIDA predict the polarity of the consumer-generated reviews significantly. However, they differ in how much variance they explain with a difference between the chi-squared values of the models of 545.69, indicating that SENTIDA is the better model.

SENTIDA and AFINN were further evaluated based on their performance in accurately categorizing the 20% of reviews that were not included in the model training as either positive or negative. The logistic regression models were applied to the untrained reviews to provide probability scores. If the probability was above 50% the review was classified as positive, whereas if it was below 50% it was classified as negative.

Through 10 tests of binary classification of Trustpilot reviews (see Table 2), SENTIDA averaged an accuracy score of 0.81 (SD = 0.01), whereas AFINN averaged an accuracy score of 0.745 (SD = 0.02). That is, on average SENTIDA correctly predicted 81% of the reviews to be either positive or negative solely based on the text, whereas AFINN on average correctly predicted 74.5 % of the reviews.

Table 2: Accuracy scores on Trustpilot reviews

	Trial										Total	
	1	2	3	4	5	6	7	8	9	10	M	SD
SENTIDA	0.80	0.81	0.83	0.81	0.81	0.80	0.80	0.81	0.81	0.81	0.81	0.01
AFINN	0.74	0.74	0.77	0.76	0.76	0.70	0.75	0.75	0.73	0.75	0.745	0.02

Note: M = Mean score, SD = Standard deviation[a]

Journal of language works

3.1.2 CORRELATION TEST

A Spearman rank-order correlation test was conducted, as this is another widely used method in validating sentiment analysis tools (Hutto & Gilbert, 2014). The correlation test was conducted on Corpus II, including 1-, 3-, and 5-star Trustpilot reviews. The decision to exclude 2- and 4-star reviews was inspired by the testing of VADER, an English sentiment tool, as Hutto & Gilbert (2014) argue that 2 and 4-star reviews are almost impossible to distinguish from 1- and 5-star reviews. The correlation was computed between the number of stars in the reviews and the mean sentiment scores of the reviews assigned using SENTIDA and AFINN.

Table 3: Correlation between model scores and Trustpilot reviews

	Correlation	p-value
SENTIDA	0.63	<.001***
AFINN	0.56	<.001***

According to the Spearman interpretation guide (Spearman Interpretation, 2018), SENTIDA (0.63) and AFINN (0.56) are both in the area of strong correlation (see table 3). However, SENTIDA's effect size is remarkably larger than AFINN's, which indicates that SENTIDA outperforms AFINN on sentiment analysis of consumer-generated reviews.

Hutto & Gilbert (2014) conducted similar correlation tests on consumer-generated reviews from *amazon.com*, which is highly comparable to the reviews on *trustpilot.dk*. The correlation test was conducted on eight different tools, with the binary Hu-liu04 (Hu & Liu, 2004) sentiment lexicon and VADER scoring the highest effect sizes of .571 and .565.

This comparison shows that both AFINN and SENTIDA would be able to compete with some of the most prominent English tools in this domain. Indeed, SENTIDA outperforms them all by a margin of at least ~.06.

3.2 Domain-specific validation corpora

The domain-specific validation corpora used in this study have been manually created by the authors in order to test SENTIDA in other domains in which sentiment analysis might be used. This test is considered to be of importance, as the authors seek to make SENTIDA a domain-independent tool that can analyse everything from poems to news articles. More specifically, this second validation ensures that SENTIDA is capable of dealing with the differences in how sentiment is expressed across different domains (Aue & Gamon, 2005).

Multiple annotators contributed in the creation of these independent test corpora that are highly comparable to SENTIDA and AFINN, in that they contain sentiment strength scored texts ranging from -5 to 5, effectively having the same metrics as the two lexicons. The test corpora are based on 45 sentiment-bearing texts from 3 different domains — that is fictional literature, political texts and social media posts. The texts were randomly chosen.

Journal of language works

Similar to the rating of the lemmas in SENTIDA, a coding scheme was employed in order to capture reliable and replicable data (see scheme below).

Six annotators (three males and three females), including the authors, were tasked with annotating the 45 pieces of texts individually using the coding scheme described above. With the data collected, the mean rating for each piece of text was computed and used to validate both SENTIDA and AFINN. The texts had a mean length of 37.3 words (SD = 15.9).

The use of manual human annotation provides a way of comparing the sentiment scores of SENTIDA and AFINN, which both rely on the bag-of-words principle, against a human reference where context is considered.

To assess the replicability and validity of the inferences made using this content, Krippendorff's alpha was used to measure agreement between raters. Using R Studio and the package *irr* (Gamer, 2015), Krippendorff's alpha was calculated to 0.901, which is acceptable to make certain conclusions using the lexicon.

3.2.1 Correlation test

Similar to the Trustpilot validation test, a Spearman rank-order correlation test was conducted, testing for the correlation between the manually annotated mean sentiment scores of the texts and the mean sentiment scores of the texts assigned using SENTIDA and AFINN.

A correlation test was conducted for each of the three categories; political texts, fictional literature and social media texts (SoMe) and across all of the categories (Table 4).

Table 4: Correlation between model scores and manually annotated texts

	Correlation coefficient		p-value	
	SENTIDA	*AFINN*	*SENTIDA*	*AFINN*
Politics	0.72	0.54	<.01**	<.05*
Fiction	0.74	0.53	<.01**	<.05*
SoMe	0.86	0.55	<.001***	<.05*
Total score	0.78	0.55	<.001***	<.001***

Note: All correlations were significant, $p < .05$

The results suggest that SENTIDA is better at predicting sentiment across all categories, outperforming AFINN with differences in correlation effect sizes of 0.18, 0.21 and 0.31 for political, fictional and social media texts, respectively. Similarly, there is a difference in the effect size of 0.23 between SENTIDA and AFINN across all categories, further consolidating the improvement SENTIDA represents in comparison to AFINN.

Journal of language works

With effect sizes above 0.5, both AFINN and SENTIDA are strongly correlated with the scores in the validation corpora, with SENTIDA showing a robust correlation having effect sizes of above 0.7 across all categories.

Furthermore, the results indicate that SENTIDA excels in the social media domain, with an effect size for the correlation of 0.86 compared to 0.72 and 0.74 for political texts and fiction, respectively.

4. Discussion

SENTIDA has proven to outperform AFINN substantially in all tests conducted, while also obtaining a higher correlation score than prominent English tools in comparable tests on consumer-generated reviews. Socher et al. (2013) elaborate on the benchmarks of the current state of sentiment analysis, arguing that for binary classification the state-of-the-art models break 80% accuracy.

In that light, SENTIDA's performance, with scores ranging from 80% to 83% in binary classification, lives up to the expectations of much more complex sentiment analysis tools. Moreover, the test done on the domain-specific text corpora provided a clear indication of SENTIDA being highly applicable across different domains.

In combining features from different existing English sentiment tools and related language studies with a stemmed wordlist constructed by the authors, SENTIDA has proved to be the best Danish tool for catching sentiment. The tool differs from the traditional bag-of-words technique by employing functions that deal with negations, adverb modifiers and exclamation marks.

In perspective, the construction of SENTIDA reflects the complexity of language, as sentiment is not only found in words themselves but also to a high degree in which context they appear. Even when taking the latter into account by incorporating a degree of context awareness, SENTIDA is far from perfect, showing how hard it is to capture sentiment fully for a sentiment analysis tool. Most interestingly, an underlying assumption of sentiment analysis is that the sentiment of a word is roughly the same for most people. That is, if language was not somewhat uniform, sentiment analysis would not be possible. However, in studies undertaken by Wilson et al. (2005) and Balahur et al. (2013), it is found that inter-annotator agreement for humans is only ~80%. This means that an objective understanding of the semantics (and sentiment) of a given text is not always shared for the population tested with the tool. In this light, a sentiment analysis tool can not be expected to surpass 80% accuracy.

Even though an elaborative discussion of whether it is possible to quantify language is beyond the scope of this paper, the authors want to emphasize that sentiment analysis tools are far from outperforming humans. Thus, SENTIDA should not be considered a replacement to classic discourse analysis but rather an entirely new method for large-scale text analysis.

This paper found that SENTIDA could be a possible solution to the observed need for new Danish sentiment analysis tools. Though the results indicate superiority in every test, there are still limitations to the tool.

Journal of language works

4.1 Methodological limitations

Although SENTIDA incorporates context sensitive measures such as negations and exclamation marks, the main methodological limitations arise from the assumptions following the bag-of-words technique. By analysing words in a text individually this technique potentially compromises pragmatic aspects of meaning as it fails to take into account the context in which the words appear. In other words, the full meaning of a text cannot be captured accurately without regards to syntax and distributional probabilities (Sahlgren, 2008). SENTIDA also has trouble interpreting homographs, i.e. words that are spelt the same but have different meanings. A similar problem is encountered with the interpretation of idioms, proverbs and adages expressing emotions metaphorically. Finally, neither sarcasm nor irony can be dealt with using SENTIDA.

When stemming, it is important to keep in mind that it relies on the assumption that all inflections of a word contains the same sentiment. This can be a problem when handling e.g. inflected adjectives such as the word *vred* 'angry' that receives the same sentiment score for all inflections, even though the inflected superlative degrees *vredere* 'angrier' and *vredeste* 'angriest' have relatively more negative connotations.

Another limitation arises in the validation of SENTIDA, in that the test corpora were made by the authors. More specifically, in the Trustpilot-corpora, the validation relies on the congruency between text and star-rating. Reviewers do not follow a fixed coding scheme for the star-rating, causing a general incongruence between text and star-rating.

For the domain-specific validation corpora, 15 texts in each domain are not enough to entirely validate a sentiment analysis tool. Furthermore, only six people rated each text, three of the six being the authors. The authors' ratings could contain some element of bias, as they also annotated the words for the SENTIDA lexicon.

4.2 Further research

In order to develop SENTIDA even further, we need to get our inspiration from the English state-of-the-art models. Today, most of these are based on machine learning approaches (Barnes, Klinger, & Walde, 2017). In these models, words are represented as functions of the context in which they occur, and by utilising machine learning algorithms, the models are then able to match these generalised functions with tokens that have similar representations. This way they can detect the exact context — and hence, sentiment rating — of a word with high accuracy (Barnes, Klinger, & Walde, 2017). This is arguably giving the models an advantage over other approaches relying on the aforementioned bag-of-words principle, which does not take context into account.

However, Hutto & Gilbert (2014) argue that machine learning models have several drawbacks, including the fact that they depend on the data sets upon which they are trained. Moreover, these models are much more computationally expensive concerning CPU processing and memory requirements, compared to models based on the bag-of-words principle, like AFINN and SENTIDA (Hutto & Gilbert, 2014). This is not just a huge disadvantage when running analyses; it also makes this type of model less convenient for the broad population.

Journal of language works

As we have devised SENTIDA with the aim of providing an accessible and easy-to-use tool, we are not going to adopt the holistic machine learning approach. In addition to this, machine learning models are yet to outperform simple models significantly (Hutto & Gilbert, 2014), which emphasises that the trade-off between computational efficiency and accuracy is not worthwhile yet, especially when keeping the results already achieved by SENTIDA in mind.

However, a hybrid model incorporating machine learning functions in a model relying on bag-of-words principle has, according to a study undertaken by researchers from UCL, shown to improve simple bag-of-words models by a margin of 5 % in accuracy scores on binary classification tasks (Kolchyna et al. 2015).

4.3 Areas of application

Looking into the possible applications of SENTIDA, a wide range of use-cases are apparent. Considering the promising results extracted in the validation, SENTIDA is to be deemed a tool that can be utilised for application in both research and business. With SENTIDA providing state-of-the-art results in the domain of consumer-generated reviews, the tool could be valuable in areas like consumer sentiment tracking and market research.

Similarly, we encourage it to be used within a broad range of language research studies, in that, remarkable results have been found using English tools comparable to SENTIDA (Hutto & Gilbert, 2014). In relation to our own academic background in Cognitive Science, SENTIDA facilitates large-scale investigations of both language and the human mind.

SENTIDA has been made freely available for download and use in RStudio. For further details see https://github.com/Guscode/Sentida.

5. Conclusion

The aim of this paper was to fill the observed gap of missing tools within the field of Danish sentiment analysis. Our new model named SENTIDA was constructed and validated on test corpora in comparison with AFINN. While both models achieved significant results, SENTIDA performed better than AFINN in all tests conducted. Likewise, a comparison of validation results with prominent English tools further solidified the competence of SENTIDA. Future development of the model is considered to include an increase in the number of words, more annotators and supplementary machine learning inspired functions.

Acknowledgements

The construction of SENTIDA would not have been possible without inspiration from Finn Årup Nielsen, who paved the way for the development of Danish sentiment analysis. Likewise, Kristian Tylén has provided invaluable support and guidance. We thank the annotators who participated in the creation of our domain-specific test corpora. Finally, the creation of SENTIDA would not have been possible without assistance from Jørg Asmussen from DSL.

Journal of language works

References

Aue, A., & Gamon, M. (2005). *Customizing Sentiment Classifiers to New Domains: A Case Study*. Retrieved from http://tiny.cc/y82h9y

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Belyaeva, J. (2013). Sentiment Analysis in the News. *ArXiv*:1309.6202 [Cs]. Retrieved from http://arxiv.org/abs/1309.6202

Barnes, J., Klinger, R., & Walde, S. S. im. (2017). Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. *ArXiv*:1709.04219 [Cs]. Retrieved from http://arxiv.org/abs/1709.04219

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Bouchet-Valat, M. (2015). SnowballC. Retrieved from https://cran.rproject.org/web/packages/SnowballC/SnowballC.pdf

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting Depression via Social Media*. Retrieved from http://tiny.cc/980e9y

Dragut, E., & Fellbaum, C. (2014). The Role of Adverbs in Sentiment Analysis. *In Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore* (1929-2014) (pp. 38–41). Baltimore, MD, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3010

Enevoldsen, K. C., & Hansen, L. (2017). Analysing Political Biases in Danish Newspapers Using Sentiment Analysis. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift*, 2(2), 87–98.

Gamer, M. (2015). *irr*. Retrieved from https://cran.r-project.org/web/packages/irr/irr.pdf

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168–177. https://doi.org/10.1145/1014052.1014073

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, 10, 216–225.

Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. ArXiv:1507.00955 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1507.00955

Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3), 411–433. https://doi.org/10.1093/hcr/30.3.411

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

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

Nielsen, F. Å. (2018). *Danish resources*. Retrieved from https://bit.ly/2NDHcbW

Pang, B., & Lee, L. (2005a). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., & Lee, L. (2005b). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 115–124). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219855

Journal of language works

Penka, D., & Zeijlstra, H. (2010). Negation and polarity: an introduction. *Natural Language & Linguistic Theory*, 28(4), 771–786. https://doi.org/10.1007/s11049-010-9114-0

RStudio Team (2015). *RStudio*: Integrated Development for R. Rstudio, Inc., Boston, MA. URL: http://www.rstudio.com

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20, 33-53.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D13-1170

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY, US: Doubleday & Co.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y. Patwardhan, S. (2005). OpinionFinder: A System for Subjectivity Analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* (pp. 34–35). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1225733.1225751

Websites

Definition of *ikke* ‘not’ retrieved on 12/12/2018, 2018, from https://ordnet.dk/ddo/ordbog?query=ikke

DSL’s 10.000 lemmas retrieved on 6/12/2018, 2018, from https://korpus.dsl.dk/e-resources/Frequently%20used%20lemmas.html

Spearman interpretation guide retrieved on 20/12/2018, 2018, from http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

The DSL list of inflections retrieved on 15/12/2018, 2018, from https://korpus.dsl.dk/e-resources/Flexikon.html

DSL www.dsl.dk