Language Resources and Evaluation (2023) 57:581–609
https://doi.org/10.1007/s10579-023-09644-5

ORIGINAL PAPER

FinnSentiment: a Finnish social media corpus for sentiment polarity annotation

Krister Lindén[1] · **Tommi Jauhiainen**[1] · **Sam Hardwick**[1]

Accepted: 1 February 2023 / Published online: 3 March 2023

Abstract

Sentiment analysis and opinion mining are essential tasks with many prominent application areas, e.g., when researching popular opinions on products or brands. Sentiments expressed in social media can be used in brand name monitoring and indicating fake news. In our survey of previous work, we note that there is no large-scale social media data set with sentiment polarity annotations for Finnish. This publication aims to remedy this shortcoming by introducing a 27,000-sentence data set annotated independently with sentiment polarity by three native annotators. We had three annotators annotate the whole data set, which provides a unique opportunity for further studies of annotator behavior over the sample annotation order. We analyze their inter-annotator agreement and provide two baselines to validate the usefulness of the data set.

Keywords Sentiment · Polarity · Social media · Finnish · Data set

✉ Krister Lindén
krister.linden@helsinki.fi

Tommi Jauhiainen
tommi.jauhiainen@helsinki.fi

Sam Hardwick
sam.hardwick@helsinki.fi

[1] University of Helsinki, Helsinki, Finland

1 Introduction

In automatic sentiment analysis of textual sources, the system aims to annotate a given text by its sentiment, which can be positive, neutral, negative, or something more fine-grained, like one of the core emotions from Plutchik's Wheel (Plutchik, 1980).

The interest and motivation for sentiment analysis behind our work at the Language Bank of Finland[1] are primarily to provide Finnish data sets automatically annotated with sentiment polarity. The need for this kind of annotation has been signaled to us by digital humanities researchers who would like to, e.g., conduct research on sentiment use in interactive conversation or the presence of communicative expressions in different sentiment contexts.

Sentiment analysis and opinion mining is an important task that has gained a lot of attention through a multitude of related shared tasks organized as part of NT-CIR (Seki et al., 2007, 2008, 2010) and SemEval workshops (Nakov et al., 2013, 2019; Rosenthal et al., 2014, 2015, 2017; Pontiki et al., 2014, 2015, 2016; Ghosh et al., 2015). Customer sentiment analysis has gained traction in commercial product and brand name monitoring (Liu et al., 2017). With the increasing use of social media, opinion mining has a prominent application in indicating fake news (Bhutani et al., 2019; Kula et al., 2020; Alonso et al., 2021).

An abundance of research has gone into sentiment analysis and data set creation in various languages, e.g., English (Bostan & Klinger, 2018), Arabic (Abdulla et al., 2014), Chinese (Ku et al., 2007), French (Apidianaki et al., 2016), and German (Clematide et al., 2012). In Sect. 2, we, in particular, look at research using or creating Finnish sentiment data, including research using the data set described in this article, first made publicly available in 2020 (Lindén et al., 2020).

During our first survey of Finnish sentiment resources in 2019, we found that an extensive data set for Finnish social media sentiment polarity research was lacking. To remedy the situation, we created a sentiment polarity annotated data set using texts from the leading Finnish social media site—*Suomi24*.[2] The various Suomi24-based data sets comprising 4.6 billion words offered by the Language Bank of Finland[3] are among the largest text corpora for the modern Finnish language, and they are actively used by Finnish digital humanities researchers (Lagus et al., 2016; Harju, 2018; Jantunen, 2018; Määttä et al., 2020). Following Boland et al. (2013) and Öhman and Kajava (2018), we decided that the sentences would be annotated without context.

We picked 100,000 random sentences from the *Suomi24* discussion forums published between 2001 and 2017 (Aller Media Ltd., 2019), which was our latest version of the data set at the time. A brief inspection found that most of the data would likely be neutral, so to make better use of the manual annotation time, we bootstrapped the procedure by creating two preliminary methods for indicating the likely

[1] https://www.kielipankki.fi/language-bank/.
[2] www.suomi24.fi.
[3] http://urn.fi/urn:nbn:fi:lb-2022011221.

sentiment of all the sentences as described in Sect. 3. We picked half of the data to be manually annotated from sentences with sentiment indications that both methods agreed on and the rest of the data from the remaining portions of the data set. The selected 27,000 sentences were divided into nine work packages outlined in Sect. 4.

We used three native Finnish speakers as manual annotators of the data set. After an initial training session, the annotators were instructed to work individually. They all received the same data packages with 3000 sentences for which to indicate positive, negative, or neutral sentiment. All three annotators completed all work packages, and in Sect. 5, we analyze the individual sentiment indications over the sample annotation order as well as their inter-annotator agreement. To validate the annotations, we also created a gold standard evaluation set by having the three authors of the paper annotate 1000 sentences picked at random from the annotation set. Therefore, we have six sentiment estimates for these sentences, which we use to evaluate the performance of the three annotators of the whole data set. We also take a closer look at some examples on which the annotators disagree.

As described in Sect. 6, based on the annotator indications for each sentence in the data set, we provide the majority vote and a derived 5-grade sentiment scale often used in shared tasks. We split the data into 20 folds for performing cross-validation. Finally, we describe the file format in which each sentence and the scores are provided.

To demonstrate the usefulness of the data set, we perform two baseline experiments with the data set in Sect. 7. We use one lexicon-based method, independent of the data set, and one neural network-based model, which we train on our data set and use cross-validation for testing. We also perform some initial analysis of where the models diverge from the human analysis and conclude the paper with a discussion and conclusion in Sect. 8.

2 Previous work

In this section, we give some suggestions for further reading on computational sentiment analysis, after which we present the previous research in Finnish language sentiment analysis and related data sets.

For a general introduction to sentiment analysis, we refer the reader to the survey by Pang and Lee (2008) in 2008. Their work was followed by Liu (2012), who gives an in-depth introduction to sentiment analysis and opinion mining and presents a comprehensive survey of all important research topics up until 2012. Feldman (2013) reviews some of the leading research questions for sentiment analysis. In 2014, Medhat et al. (2014) surveyed the algorithms and applications for sentiment analysis. They intended to update the earlier work and give newcomers a panoramic view of the field. They also categorize the available benchmark data sets at the time. Later, also Ravi and Ravi (2015) give a survey on opinion mining and sentiment analysis and list publicly available data sets known to them. Later additions to surveys concerning sentiment analysis have been made by Giachanou and Crestani (2016), who discuss sentiment analysis for Twitter, and Zhang et al. (2018), who survey deep learning techniques used in sentiment analysis. Mäntylä

et al. (2018) present a computer-assisted review of the evolution of sentiment analysis by analyzing 6996 papers from Scopus.

The earliest work we have identified that discusses automatic sentiment analysis for the Finnish language is Tiia Leuhu's Master's Thesis (Leuhu, 2014). Tweets in Finnish were manually annotated so that the collection consisted of 700 tweets in each of the three categories: positive, neutral, or negative. Using 10% of the data for testing, she evaluated three machine learning algorithms: k-nearest neighbor, multinomial naïve Bayes, and random forest. Naïve Bayes proved to be the best algorithm for sentiment classification, attaining an accuracy of 0.84. The annotated data set was not published.

Paavola and Jalonen (2015) used sentiment analysis in order to detect trolling behavior in tweets in Finnish during the 2014 Ukrainian crisis. They used a social media analysis tool developed in the NEMO project to detect the polarity (positive-neutral-negative) of the messages. The tool uses pre-defined positive and negative words and emoticons together with a decision tree and logistic regression. The work was continued by Paavola et al. (2016a, b), who analyzed Finnish tweets during the Syrian refugee crisis in order to detect bots. The tool is not currently available.

Öhman et al. (2016) used the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013) to study the preservation of sentiments in translation in the Opensubtitles parallel corpus of movie subtitles (Lison & Tiedemann, 2016) as well as the Europarl corpus in OPUS (Tiedemann, 2012). The word-emotion association lexicon was used to label sentences with one of the eight core emotions from Plutchik's wheel (Plutchik, 1980) in addition to being generally negative or positive. The language pairs investigated were English—Finnish, English—Swedish, and Spanish—Portuguese. Using manually annotated sentences, they found that the Spanish—Portuguese pair has a higher cross-language agreement than the other two pairs.

Jussila et al. (2017) investigated the reliability of two sentiment analysis tools for Finnish when compared with human evaluators. The two analysis tools were the SentiStrength (Thelwall et al., 2010, 2012) and the Nemo Sentiment and Data Analyzer (Paavola & Jalonen, 2015). The Nemo Sentiment and Data Analyzer tool can also be used to collect tweets, and it was used to collect a set of 509 tweets in Finnish. Two human annotators independently classified each tweet as positive, negative, or neutral. The Nemo Sentiment Analyzer can use one out of two separate algorithms to analyze sentiments: logistic regression and random forest. The SentiStrength returns the strength of positive and negative sentiment of the text on a scale from one to five. The values given by the three algorithms were used to classify the tweets as positive, negative, or neutral. The automatic classifications were then compared with the classifications of two human annotators. They used Krippendorff's alpha (Krippendorff, 2011) for evaluating the inter-annotator agreement and reliability of the annotations. The annotated data set was not published.

Kaustinen (2018) used a Finnish data set with 14,332 movie reviews rated from 1 to 10. The data was gathered from leffatykki.com in November 2017. He investigated the effect linguistic differences between English and Finnish have on sentiment analysis.

In his Master's Thesis, Nukarinen (2018) used deep learning, Long Short-Term Memory (LSTM) recurrent neural networks, in experimenting with sentiment analysis in Finnish. For his experiments, he gathered over 50,000 product reviews from http://www.verkkokauppa.com. When classifying into categories from one to five, his classifier achieved an overall accuracy of 53.6%. He did not publish the data set.

Öhman and Kajava (2018) and Öhman et al. (2018) introduce a web-based annotation tool called *Sentimentator*.[4] Sentimentator uses a ten-dimensional model based on Plutchik's core emotions. Annotating sentences using a ten-dimensional scheme requires more reflection from the annotator than simply tagging the sentence as positive, negative, or neutral. The authors set out to solve this by gamifying the process. In order to avoid domain bias, they set out to annotate the texts at the sentence level without a larger context, as also suggested by Boland et al. (2013). They used the Opensubtitles data set from OPUS with an initial focus on English and Finnish.

Kajava (2018) and Kajava et al. (2020) investigated sentiment preservation in translations and transfer learning. Continuing the utilization of the Opensubtitles corpus, they used English sentences as the source and their Finnish, French, and Italian translations as targets. Each sentence was labeled with one of Plutchik's core emotions using the Sentimentator annotation tool (Öhman & Kajava, 2018). Once labeled, the English sentences were exported from Sentimentator and manually revised by a native English speaker who removed ambiguous or neutral sentences from the data set. The translations of the remaining sentences in Finnish, French, and Italian were similarly annotated by competent speakers, two for each language, and labeled with precisely one of the core emotions according to the speakers' judgment. The categorization of each sentence as negative or positive was then derived from these labels. In total, the data set consists of 6427 sentences for each language. Cohen's Kappa coefficient was used as a measure for inter-annotator reliability (Cohen, 1960). The sentiment preservation accuracy between English and translated sentences ranged from 0.82 for Italian to 0.86 for Finnish, indicating that sentiment is relatively well preserved in translations. Kajava (2018) also created an evaluation data set with training and testing partitions and evaluated four machine learning classification algorithms: multilayer perceptron (MLP), multinomial naïve Bayes (MNB), support vector machine (LinearSVC), and maximum entropy (MaxEnt).[5] Depending on the language, the best classification results were given by MNB, LinearSVC, or MaxEnt classifiers.

Einolander (2019) analyzed textual customer feedback from Telia Finland. Several classification models were compared, and a deep learning model utilizing LSTM networks performed the best.

Hämäläinen and Alnajjar (2019) trained a sentiment prediction model for English (Feng & Wan, 2019) using sentiment annotated data from the OpeNER project (y Montse Cuadros y Seán Gaines y German Rigau, 2013). Then they mapped pretrained fasttext models for English and Finnish into a common space. The resulting

[4] https://github.com/Helsinki-NLP/sentimentator.

[5] The data is available at https://github.com/cynarr/MA-thesis/tree/master/data-raw.

sentiment prediction model for Finnish is available as part of the FinMeter library (Hämäläinen & Alnajjar, 2021).[6]

Vankka et al. (2019) implemented polarity lexicons for Finnish. They used reviews written in Finnish from the Trustpilot and TripAdvisor websites. The reviews were rated with values from 1 to 5. They created a hybrid algorithm using the polarity lexicons together with word embeddings. They found that using the headlines of the reviews instead of their content was less noisy as the content often describes both the negative and positive sides of the reviewed item. The corpus they used is not currently available. Later, Vankka et al. (2021) proposed using cross-lingual projection in sentiment analysis when developing a framework for sentiment analysis and clustering of Finnish and Russian tweets. They used word-embedding vectors to learn a translation matrix to project Finnish tweets to the Russian target space. They compared the accuracy of these projections in sentiment classification with their earlier monolingual classifier (Vankka et al., 2019) and found that the cross-lingual approach performed worse.

Kuutila et al. (2020) used sentiment analysis to determine software developer productivity. For sentiment analysis, they translated lexicons used for measuring arousal and valence from English to Finnish. In addition to these lexicons, they used lists of emoticons that had been manually classified to Plutchik's basic emotions. The translated lexicons and the list of emoticons are available at GitHub.[7]

Öhman (2020) presents a continuation of the work using Sentimentator and the OPUS Movie Subtitle parallel corpus to annotate individual subtitle lines with Plutchik's core sentiments. She primarily focuses on describing and evaluating the annotation process in detail. The result of the annotation work was over 56,000 annotated sentences in Finnish, Swedish, or English by roughly 100 separate annotators. Öhman et al. (2020) published the XED data set with 25,000 Finnish and 30,000 English sentences annotated with Plutchik's core emotions.[8] The XED data set is the largest release from the Helsinki-based research group so far, continuing the work with Sentimentator (Öhman & Kajava, 2018) and open movie subtitle data from OPUS (Tiedemann, 2012). In addition to Finnish and English, the release includes projected annotations for 30 other languages. Öhman (2021) prepared manually verified versions of Finnish sentiment and emotion lexicons originally published by Mohammad and Turney (2013). The resulting lexicons, Sentiment and Emotion Lexicon for Finnish (SELF) and Finnish Emotion Intensity Lexicon (FEIL), are available for download from Github.[9]

In his Master's Thesis, Karttunen (2021) studied the relationship between investor sentiment and stock prices. He created several Bayesian classifiers using the first version of the data set described in this article (Lindén et al., 2020). His research did not find any connection between social media post sentiments and stock prices.

[6] https://github.com/mikahama/finmeter.
[7] https://github.com/M3SOulu/semotion2020.
[8] https://github.com/Helsinki-NLP/XED.
[9] https://github.com/Helsinki-NLP/SELF-FEIL.

Rautiainen and Luoma-aho (2021) studied the links between social media sentiment and the perception of public companies' stakeholders as well as their financial statement information. For sentiment analysis, they used the *M*-adaptive program by M-brain.[10]

Hellström (2022) studied aspect-based sentiment analysis for the Finnish language in his Master's Thesis. He proposes two solutions, both of which use gradient harmonized and cascaded labeling (Luo et al., 2020) together with FinBERT (Virtanen et al., 2019) for aspect polarity classification. For training and testing, he used 1673 sentences from product reviews from Verkkokauppa.com that he had manually annotated.

According to our recent review, in 2022, in addition to our own 27,000-sentence data set based on social media, the only publicly available Finnish language data sets with manual sentiment annotations are the 6427 sentences published by Kajava (2018) and the 25,000 sentences by Öhman et al. (2020) based on movie subtitles.

3 Preliminary sentiment annotations

We implemented a CNN sentence classifier (Kim, 2014) for classifying texts for sentiment polarity before our current work. We trained this architecture on two data sets: a collection of product reviews scraped from online web stores and sentences from the *Suomi24* corpus containing emoticons.

External users initially tested these tools, but their reliability was deemed relatively low. For some tasks like psychological priming experiments, the analyzer based on product reviews was felt to correlate better with human evaluations. These experiences led us to embark on a more extensive manual effort to annotate social media sentences with sentiment polarity.

However, despite some social media discussions being inflamed, much of the text is still relatively neutral. To use the human annotation effort efficiently, we decided that the preliminary sentiment analyzers could be used to weed out some of the neutral sentences. Doing this would raise the odds that there was at least a considerable number of sentences with sentiment polarity in the data to be given to the human annotators.

We pre-trained word embeddings for the model with +word2vec+ (Mikolov et al., 2013) using the entire *Suomi24* corpus by Aller Media Ltd. (2019).

3.1 Product review-based annotator

The product reviews contained a review text and a star rating, from 1 to 5 stars, reflecting total product satisfaction. We mapped this rating to a three-way sentiment classification by assigning 3 as neutral, < 3 as negative, and > 3 as positive.

[10] https://www.m-brain.com.

Table 1 Distribution of pre-selected sentences

		Smiley		
		POS	NEUTR	NEG
Product review	POS	4861	24,984	895
	NEUTR	3007	18,914	1891
	NEG	4494	35,274	5680

3.2 Smiley-based annotator

We took the intentionally naïve approach of directly taking a very limited interpretation of smileys as cues of sentiment in sentences. Those texts containing only positive smileys were assessed as positive, texts containing only negative smileys were assessed as negative, and texts containing neither were assessed as neutral. Texts containing both positive and negative smileys were entirely discarded. Emoticons were used as distant supervision similar to Read (2005), Pak and Paroubek (2010), and Abdul-Mageed and Ungar (2017).

4 Corpus

The original corpus consists of sentences from the social media site *Suomi24*,[11] which is available as a corpus through the Language Bank of Finland. From this corpus, we randomly selected sentences and pre-annotated them with the pre-annotators for screening purposes. Based on the pre-annotations, we composed a corpus that was likely to have a higher proportion of non-neutral sentences annotated by human annotators for sentiment polarity.

4.1 Suomi24

Suomi24 is one of Finland's largest social media sites. Its discussion forums, which comprise hundreds of subforums dedicated to discussion and personal advertisements, have been popular since the early 2000s. Suomi24's operators have co-operated in data sharing with the Language Bank of Finland, resulting in a 4.6 gigaword corpus spanning the years 2001–2020 being available to researchers. Largely thanks to that effort, it has been used as a data source for research in communications and sociology as well as language technology.

Each message is posted to a particular subforum and is part of a particular discussion thread. Most messages are replies to other messages. Sentiment analysis from text is a highly contextual task. When analyzing the sentiment of a single sentence, as the present work attempts, multiple levels of context must be inferred or guessed:

[11] www.suomi24.fi.

Table 2 Distribution of selected sentences

		Smiley		
		POS	NEUTR	NEG
Product review	POS	4500	4797	170
	NEUTR	573	4500	356
	NEG	869	6735	4500

the rest of the message, the other messages in the thread, and the local messaging culture of a given subforum. The problem is tractable, and the corpus provides a richly challenging setting for sentiment analysis.

4.2 Text selection procedure

First, we built a pre-selection corpus of 100, 000 random sentences from the *Suomi24* corpus [data set release 2017H2 by Aller Media Ltd. (2019)], without filtering on the basis of length or other criteria.

We pre-evaluated our sample with our two automatic annotators, *Product review* and *Smiley*, and selected the sentences for human evaluation based on this pre-evaluation. As a result, the present corpus cannot be considered directly statistically representative of sentiment of the Suomi24 corpus but has, in effect, been enriched for sentiment to avoid annotator fatigue as far as possible due to an abundance of sentences with a neutral sentiment. However, a representative sample can be reconstructed based on the documented selection procedure.

The sentences in the pre-selected corpus were classified by the automated annotators as shown in Table 1.

The automated pre-evaluation annotators completely agreed on 29,455 sentences, slightly disagreed (one was neutral and the other was not) on 65,156 sentences, and strongly disagreed on 5389 sentences.

This pre-selection corpus was then divided into four categories, which were used for selection into the final corpus in desired proportions. Well aware that annotating may sometimes be a time-consuming task, we also wanted to divide the work into work packages for the human annotators to let them feel that they had made visible progress when a work package had been completed. In each work package of 3000 sentences, we included sentences evaluated by both our automated pre-evaluation annotators, of which

- 500 had an agreed-on positive sentiment,
- 500 had an agreed-on neutral sentiment,
- 500 had an agreed-on negative sentiment, and
- 1500 on which the automated annotators disagreed

As a result, the sentiment corpus of 27,000 sentences had potentially positive, neutral, and negative evaluations with 4500 sentences each and 13,500 sentences with

Table 3 Distribution of annotations

Annotator	Positive	Neutral	Negative	Pos-neg ratio
A	4576 (17.0%)	15,927 (59.0%)	6497 (24.1%)	70.4%
B	3267 (12.1%)	18,459 (68.4%)	5274 (19.5%)	61.9%
C	2118 (7.8%)	22,954 (85.0%)	1928 (7.1%)	109.9%
Average	3320 (12.3%)	19,113 (70.8%)	4566 (16.9%)	72.2%

potentially no clear sentiment polarity. The corpus with potentially enriched polarity data had the distribution shown in Table 2.

The 27,000 sentences comprised a total of 346,937 tokens and 2,052,900 (Unicode) characters, which is an average of 12.8 tokens per sentence and 76 characters per sentence.

4.3 Annotators and annotation schema

The annotators were students of language technology at the University of Helsinki. They were, however, unaccustomed to sentiment annotation, and we determined that in the interest of being able to obtain a sufficiently large corpus in a reasonable amount of time, it would be best to perform only a three-way annotation: positive, negative, and neutral.

4.4 Annotation process

We assigned the nine work packages of 3000 sentences to each of our annotators. Each package contained the same polarity distribution of sentences from our pre-selection categories. However, the sentences within each package were randomly shuffled, i.e., the sentences from each category did not appear consecutively. The annotators were blind to the pre-selection, i.e., they did not know the annotation scores. In this way, the annotation should not be biased by appearing in a long string of similarly annotated sentences or by a previous estimate. The work packages given to each annotator were identical.

After a brief initial meeting, the annotators worked independently of each other. They used a spreadsheet program to input their single-character annotation in column A for the sentence in column B.

There was no schedule set except for a final deadline, and the bulk of the annotations was performed closer to the deadline than at the beginning of the project.

We invited the annotators to a briefing to kick off the annotation task. We described the task and advised the annotators that human agreement in this task usually is in the 70% range. We explained that since the sentences were being presented out of context, it would not always be possible to judge the intended sentiment accurately, but they should avoid *overthinking* and make a quick judgment call as to whether the sentiment was either explicitly present or overwhelmingly likely in

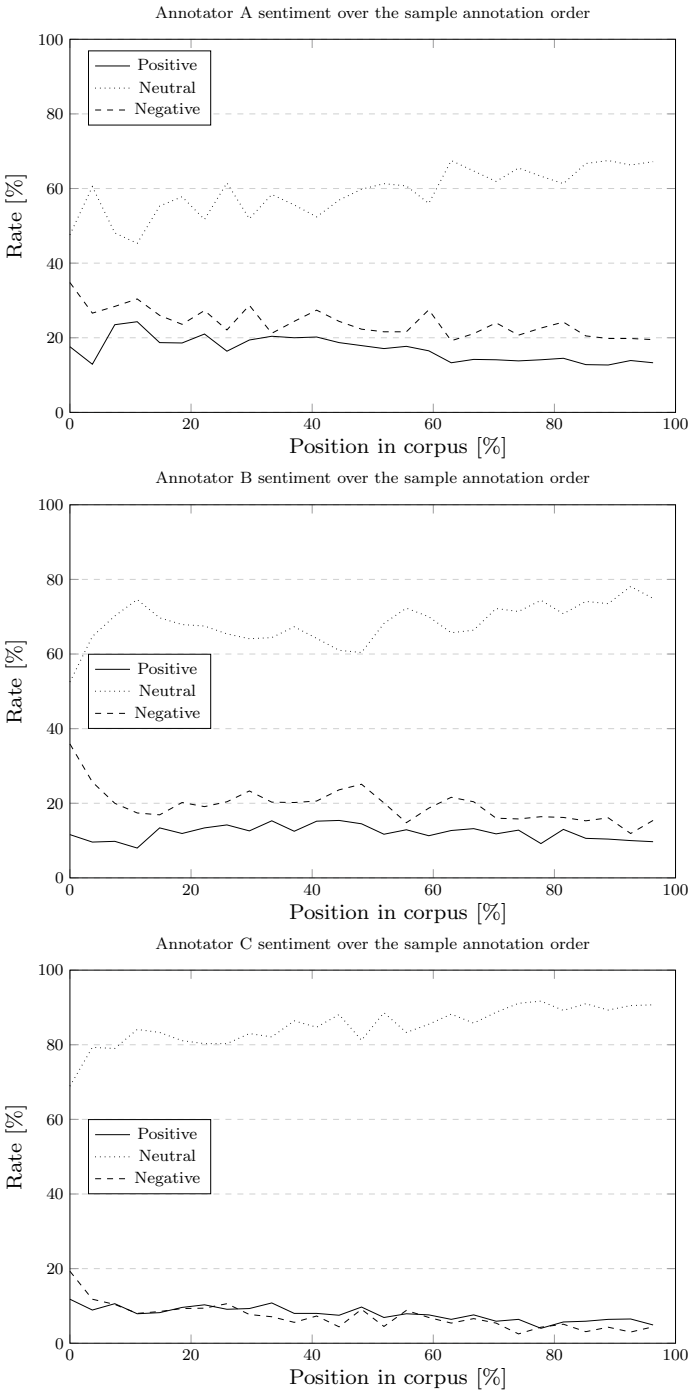

Fig. 1 Annotator-assigned sentiment over the sample annotation order

Table 4 Coincidence matrix of annotator pairs

		A		
		POS	NEUTR	NEG
B	POS	2651	552	64
	NEUTR	1621	14,109	2729
	NEG	304	1266	3704
		A		
		POS	NEUTR	NEG
C	POS	1868	133	177
	NEUTR	2631	15,571	4752
	NEG	77	223	1628
		B		
		POS	NEUTR	NEG
C	POS	1619	310	189
	NEUTR	1641	17,779	3534
	NEG	7	370	1551

context. In cases of conflicting sentiment, we advised selecting a positive or negative sentiment if one was clearly the dominant one and otherwise selecting a neutral sentiment. No written instructions were given.

After some discussion, the annotators did a trial run of 100 sentences to ensure they had some shared understanding of the task. We went over these annotations together. After this meeting, the annotators did not discuss their annotations with each other.

5 Analysis of the annotations

To see how well the annotation schema was adhered to and how sentiment perception may vary between individuals, we look at the overall distribution of sentiment ratings. Then we overview the annotations by an individual annotator for each sentence in the corpus over the sample annotation order. We also perform a second validation annotation by different annotations of a subset of the corpus and finally look at some examples to understand annotator agreement and disagreement.

5.1 Distribution of annotations

In Table 3, we see the corpus distribution of perceived sentence polarity for each annotator. Both annotators A and B find more negative than positive statements, whereas annotator C finds a roughly equal amount. In Fig. 1, we see a tendency that

Table 5 Annotator agreements

Annotators	Agreement	Strong disagreement	Krippendorff's alpha
A and B	20,464 (75.8%)	368 (1.4%)	0.54
A and C	19,067 (70.6%)	194 (0.7%)	0.34
B and C	20.949 (77.6%)	196 (0.7%)	0.44
A, B and C	16,866 (62.5%)	505 (1.9%)	0.44

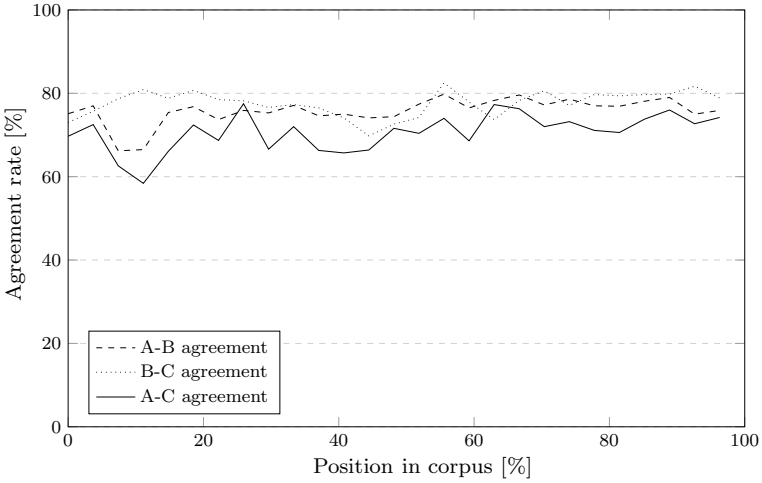

Fig. 2 Annotator agreement over the sample annotation order

is consistent for all three annotators over time, i.e., the number of statements perceived to be neutral grows towards the end of the task, but their ratio of positive vs. negative remains essentially the same.

5.2 Inter-annotator agreement

We computed *agreement*, i.e., how often annotators made the same annotation, *strong disagreement*, i.e., how often one annotator annotated a sentence as positive and another as negative, and *Krippendorff's alpha* (Krippendorff, 2011).

Krippendorff's alpha is convenient because it generalizes to scoring the agreement between more than two annotators. Because the human annotators had the task of making a categorical judgment, rather than using a finer scale, we have used the nominal level of measurement in calculating Krippendorff's alpha, meaning that all disagreements have the same weight whether between negative and neutral or between negative and positive.

In Table 4, we see how the annotators agreed on the data set level.

Table 6 Verification annotation

Annotator	Average error	# of disagreements with majority
A	0.25	167
B	0.22	125
C	0.27	189
D	0.30	281
E	0.22	171
F	0.24	198

In Table 5, we calculated the agreement, strong agreement, and Krippendorf's alpha between the annotators on the data set level.

Out of the 505 instances of strong disagreement among human annotators, 252 were cases where each of the three possible annotations was selected by an annotator, meaning that in these cases, there was no majority opinion.

5.3 Inter-annotator agreement timeline

Figure 2 shows how the inter-annotator agreement developed over time. When more than half of the corpus had been annotated, there seemed to be more agreement between the annotators, whereas their agreement on sentence polarity was less in the initial part of the corpus.

5.4 A verification annotation and annotator reliability

We chose 1000 random sentences from the corpus for annotation by a separate group of three annotators, namely the authors of this article. We summarize the result of this verification annotation in Table 6. "Average error" means the average difference between an annotator's evaluation on the scale $\{-1, 0, 1\}$ and the average of all annotators on the same sentence.

We also annotated the 505 sentences that the annotators had strong disagreements about, i.e., those that had both positive and negative annotations. We present some conclusions from this exercise in Sect. 5.5.

5.5 Some example annotations

To illustrate the content of the corpus and the task that the annotators were faced with, we provide some examples from the corpus of some cases we consider indicative of non-obvious choices made by the annotators.

All human annotators tended to agree on a positive sentiment when the sentence contained only a positive assessment of something, whether the commentator's mood, some topic of conversation, or another commentator, even if the sentiment was only a minor part of the comment:

"no mielestäni kuulostat mielenkiintoiselta, olen itse samankaltaisista asioista kiinnostunut nainen, en pidä baareista, kesällä kun on vapaata olen mieluummin puistossa tai rannalla, mutta puistoista lähden sitten siinä vaiheessa kun muut tulevat sinne ryyppäämään."

Well, I think you sound interesting, I'm a woman interested in similar things, I don't like bars, in the summer when I have some spare time I prefer to spend time in a park or on the beach, but I leave the parks when other people get there to booze.

A pos, B pos, C pos

Annotators also agreed on the positive sentiment of sentences in cases where there was a clear and unambiguous expression of tone by using words indicating politeness or smiley faces. E.g.:

"Kiitos kaikille vastaajille!"
Thanks to everyone who replied!

A pos, B pos, C pos

Here is a positively annotated case with no explicitly positive content, but which is conciliatory in tone:

"Itse asiassa pystymetsäläiset ja kruunuhakalaiset on ihan yhtä hyvää jengiä, ei tee tiukkaa."
Actually people from the countryside and the city are just as good people, no doubt.

A pos, B pos, C pos

This direct statement of the commentator's satisfaction with his situation was annotated as positive:

"Joo kyllä itse olen ihan tyytyväinen palkkaani."
Yeah, I'm quite satisfied with my salary.

A pos, B pos, C pos

Negative mood, even when not directly indicating sentiment, was annotated as negative, as in the following example, which all human annotators marked as negative:

"Nuku hyvin, Viivuska :'(♡"
Sleep well, Viivuska :'(♡

A neg, B neg, C neg

This comment indicating that an argument is taking place was annotated as negative:

"Missä kohtaa olen sinua nimitellyt?"

Where exactly did I call you names?

A neg, B neg, C neg

Some annotations, such as this negative one, require considerable knowledge about the world to interpret and assess:

"Vihreä puolue ei ole edustanut vihreitä arvoja enää ainakaan puoleen vuosikymmeneen."
The green party hasn't represented green values for at least half a decade.

A neg, B neg, C neg

Annotators selected differing annotations, especially in cases where multiple sentiments were expressed, as in this case where each positive, negative, and neutral sentiment was selected:

"Haastattelu meni tosi hyvin ja portfolioon olen panostanut paljon mutta en siltikään usko että pääsen koska en ole käynyt lukiota eikä ne mielellään ota meikäläisiä :/"
The interview went really well and I put a lot of work into my portfolio but I still don't think I'll get in because I didn't go to secondary school and they don't like to choose people like us :/

A pos, B neg, C neu, D neg, E neg, F neg

Sometimes the overall sentiment of a sentence is dependent on which category the person writing the text belongs to. In the following, if the writer is a "fat chick" herself, this could be considered very positive. Otherwise, maybe not so much.

"Joten läskit muijat: YLÖS,ULOS ja LENKILLE, ei kuitenkaan makkaralenkille ;););)"
So fat chicks: get UP,OUT and GOING, though not for lenkki-sausages ;););)

A neg, B neg, C pos, D neg, E neu, F neu

"On ja samaan klassiin kuuluu noi Muslimit ihan samalla lailla :))"
It is and those Muslims belong to the same class in the same manner :))

A pos, B neg, C pos, D neg, E neu, F neu

"Hienoissa kantimissa kuitenkin tämän hyvinvointivaltion asiat tällä puolella :)"
This welfare state is doing just great as far as this is concerned :)

A pos, B neg, C pos, D pos, E neu, F pos

The use of unusual idioms like "hienoissa kantimissa" (*in great condition*) together with more official-sounding words like "hyvinvointivaltio" (*welfare state*) may be a

Table 7 Majority vote distribution

	#
Positive	3066 (11.4%)
Neutral	19,825 (73.4%)
Negative	4109 (15.2%)

sign of sarcasm. In addition, using the word "kantimissa" might have a negative sentiment in itself as it is roughly ten times more likely to be found preceded by a negative pre-modifier like *weak* or *bad* than a positive one in the Internet texts (City Digital Group, 2021; Ylilauta, 2015) available at the Language Bank of Finland.

5.5.1 Strongly ambiguous examples

Of the 505 sentences that were evaluated both positively and negatively, we present some that still had no majority, even after the authors annotated them.

"ARMAHTAKAA rakkaat kommentoijat, en pysty enää nykyään lukemaan kunnolla tätä palstaa koska lähes aina kuolen nauruun..."
HAVE MERCY dear commenters, I can't read this forum anymore properly because I almost always die of laughter...

A neg, B neg, C pos, D neg, E pos, F pos

"No mut kyllä mä silti haluan uskoa aurinkoisempaan huomiseen mut neuvot on kyllä aika vähissä..halusin vain ilmoittaa sinulle ettet toki ole ainoa ongelmiesi kanssa."
Well, but I still want to believe in a brighter tomorrow, but the options are rather few..I just wanted to tell you that you're certainly not alone with your problems.

A pos, B neg, C pos, D pos, E neg, F pos

When the post is either gleeful or hopeful, the overall tone is read differently by different annotators.

"Teen kaiken firman tuloksen puolesta!"
I'll do everything for the company's profit!

A pos, B neg, C pos, D pos, E neg, F neg

6 Data set

Based on the annotations we had obtained, we proceeded to create two gold standard data sets of the annotations. One took the majority vote of the annotations, and the other derived a 5-grade scale often used in shared tasks.

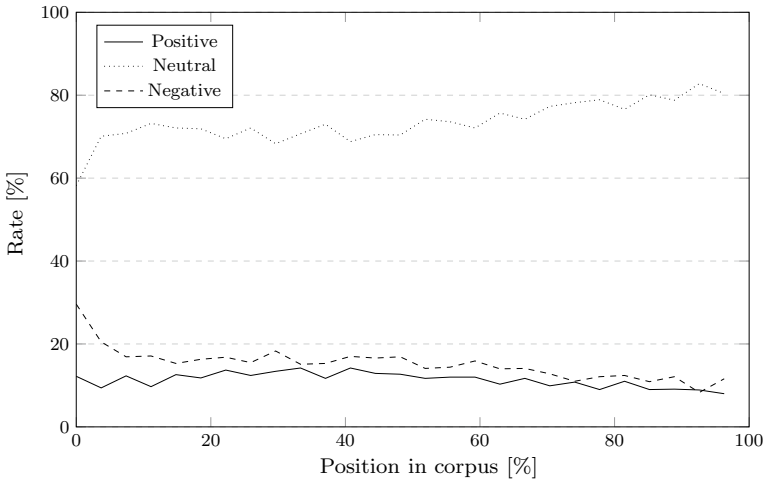

Fig. 3 Majority vote sentiment over the sample annotation order

Table 8 Derived score distribution

Sum of evaluations in this corpus	Derived category	Number in corpus
−3	1	1387 (5.1%)
−2 or −1	2	6422 (23.8%)
0	3	14,195 (52.6%)
1 or 2	4	3460 (12.8%)
3	5	1536 (5.7%)

6.1 Majority vote

The easiest way to form a gold standard of a polarity annotated corpus of three annotators is to take the majority polarity of the manual annotators and give a neutral reading for cases where all the annotators disagree. The distribution of the majority vote is shown in Table 7, and the distribution over the sample annotation order is shown in Fig. 3. The increase in the share of neutral sentiment over the sample annotation order suggests either annotator fatigue or annotator quality improvement. According to the verification annotation by the authors, there seems to be no indication of the material being more neutral towards the end of the data set, so we are inclined to believe that this reflects annotator fatigue, i.e., the annotators adapt to the general style of the data set and mark more samples as neutral.

Table 9 Data set format

Column #	Column name	Range/data type
1	A sentiment	[−1, 1]
2	B sentiment	[−1, 1]
3	C sentiment	[−1, 1]
4	Majority value	[−1, 1]
5	Derived value	[1, 5]
6	Pre-annotated sentiment smiley	[−1, 1]
7	Pre-annotated sentiment product review	[−1, 1]
8	Split #	[1, 20]
9	Batch #	[1, 9]
10	Index in original corpus	Filename & sentence id
11	Sentence text	Raw string

6.2 Derived categories (1–5)

We also report sentiment on a 1–5 scale for each sentence for compatibility with other sources. With +1 signifying positive sentiment, −1 signifying negative sentiment, and 0 signifying neutral sentiment by a human annotator, we sum the three human scores and map them to the 1–5 scale according to Table 8. By construction, this scale seemingly reflects unambiguity or clarity of sentiment rather than the strength of sentiment. To assess whether it also reflects the strength of sentiment, we chose a random 100-sentence sample from each of the five categories given by the scale. The authors independently analyzed the sentiment of each sentence on a 1–5 scale. The majority result was that there appears to be a strong correlation between the categories and the sentiment strength; the average 1–5 scale sentiment score for sentences in each derived vote category was, respectively, 1.7, 2.1, 2.8, 3.4, and 4.1. Presumably, a strong sentiment is most often also a clearly indicated sentiment and vice versa.

6.3 Splitting the data set

We created a split of the data to enable a 20-fold cross-validation corresponding to randomly shuffling the sentences and splitting them into 20 equally-sized portions. In each validation run, a different 5% section can be used for testing, another for development, and the remaining 90% as training data. In the gold standard data file, we indicate in which split each sentence ended up for comparability with our test results. If cross-validation with fewer splits is preferred, one can combine several splits for testing and development and the remaining portions for training.

6.4 File format

The corpus is available in a UTF-8 encoded TSV (tab-separated values) file with columns as indicated in Table 9. In the table, *split* refers to the cross-validation split

Table 10 Sentistrength conversion to sentiment polarity

Score	Sentiment polarity
< 0	Negative
0	Neutral
> 0	Positive

Table 11 SentiStrength polarity distribution and evaluation

Annotator	Positive	Neutral	Negative	Pos-neg ratio
SentiStrength	7163 (26.5%)	17,586 (65.1%)	2251 (8.3%)	3.18
Annotator	Accuracy	Positive F1	Negative F1	Neutral F1
SentiStrength	63.9%	0.368	0.770	0.306

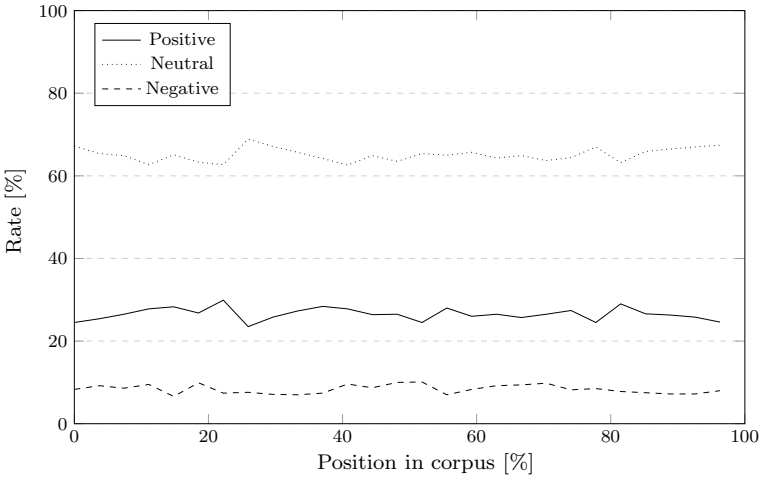

Fig. 4 SentiStrength sentiment polarity over sample annotation order

to which a sentence belongs, and *batch* to the work package the sentence belongs to. Indexes to the original corpus are strings consisting of a filename, like +comments2008c.vrt+, a space character, and a sentence id number in the file.

7 Initial experiments with the data set

To evaluate the usefulness of the gold standard data set with a majority vote and the derived scores of the manually annotated corpus, we tested the data set with SentiStrength (Thelwall et al., 2010), which is a lexicon-based sentiment analysis program using word lists for various languages. It also has word lists for Finnish. To evaluate the performance of our baseline CNN architecture on different splits of the data set,

Table 12 SentiStrength conversion to derived score

Score	Derived score
$-4 \leq score \leq -3$	1
$-2 \leq score \leq -1$	2
$score = 0$	3
$1 \leq score \leq 2$	4
$3 \leq score \leq 4$	5

Table 13 SentiStrength derived score distribution and evaluation

Annotator	1	2	3	4	5
Senti-Strength	368 (1.4%)	1883 (7.0%)	17,586 (65.1%)	7015 (26.0%)	148 (0.55%)
Annotator	Accuracy	F1 1	F1 2	F1 3	F1 4	F1 5
SentiStrength	49.7%	0.108	0.207	0.687	0.285	0.072

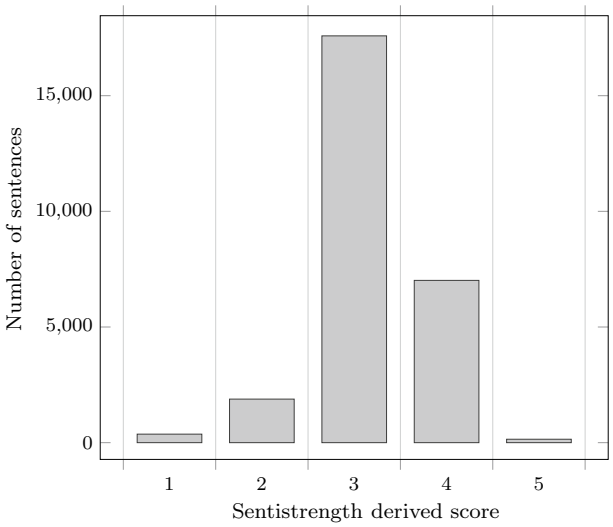

Fig. 5 SentiStrength derived score distribution

we used the 20-fold cross-validation split to train separate models. The model is as described in (Kim, 2014), Except for adding kernels of size 2 (Kim used only sizes 3, 4, and 5).

7.1 Evaluation measures

As evaluation measures for the baseline methods, we use accuracy and F1.

Table 14 CNN polarity distribution and evaluation

Annotator	Positive	Neutral	Negative	Pos-neg ratio
CNN 3-class classifier	2559 (9.5% ±1.8%)	21,668 (80.3% ± 3.5%)	2773 (10.3% ± 3.2%)	0.92
Annotators	Accuracy	F1 positive	F1 neutral	F1 negative
CNN 3-class classifier	79.0%	0.668	0.870	0.411

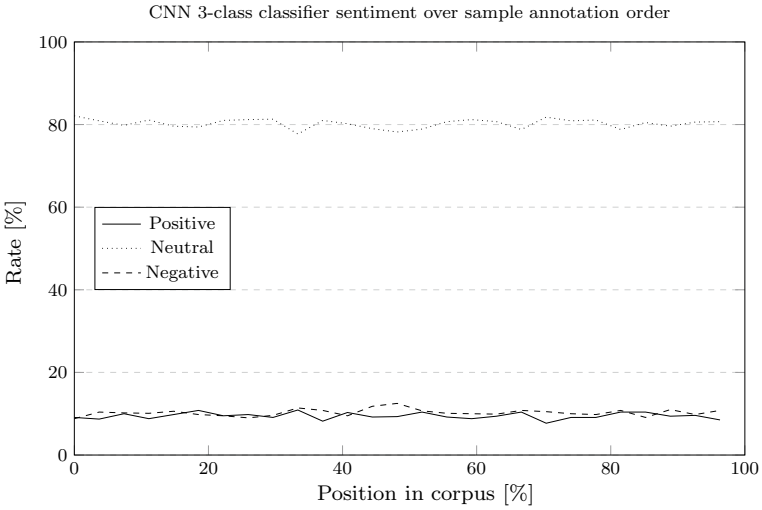

Fig. 6 CNN sentiment polarity over sample annotation order

7.2 Testing a lexicon-based model

We obtained the SentiStrength (Thelwall et al., 2010) rule-based sentiment analysis program and word lists for analyzing Finnish texts from its authors. It provides both a positive and negative sentiment score for each sentence between 1 and 5. Taking $score = score_{positive} - score_{negative}$, we convert between scales to be compatible with the majority vote and the derived score. The conversion to sentiment polarity is shown in Table 10.

We obtained the results displayed in Table 11 and illustrated in Fig. 4.

The conversion of SentiStrength scores to derived scores is shown in Table 12.

We obtained the results displayed in Table 13 and illustrated in Fig. 5. The mean absolute error averaged over the data set was 0.64, and the standard deviation of the absolute error was 0.73.

Table 15 CNN derived distribution and evaluation

Annotator	1	2	3	4	5
CNN architecture	483 (1.8% ± 1.1%)	6425 (23.8% ± 7.7%)	15,493 (57.4% ± 4.5%)	3744 (13.9% ± 5.3%)	855 (3.2% ± 1.5%)

Annotator	Accuracy	F1 1	F1 2	F1 3	F1 4	F1 5
CNN derived score	53.0%	0.188	0.395	0.658	0.285	0.072

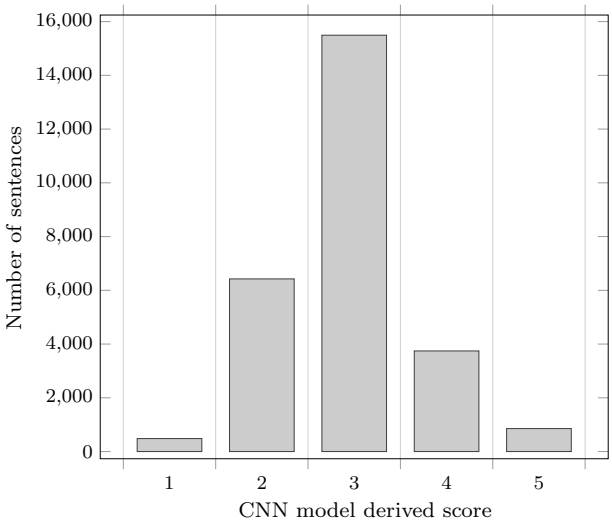

Fig. 7 CNN derived score distribution

7.3 A CNN baseline model

To evaluate the average performance of the baseline CNN architecture on the data set, we used the 20-fold cross-validation split of the data set to train 20 different CNN models.

In the first model, we used sentences belonging to splits 1 for testing, and 2 for development and 3–20 for training. We gradually shifted the testing and development splits over the whole corpus until we trained 20 models.

We trained each CNN model with the same architecture as in the preliminary annotations, fitting a mean square error function and obtaining the following results when the regression output value has been scaled to the range [1, 5] and rounded to the nearest integer.

Using the human majority vote in the gold standard data set as training and test data, we obtained the following results for the 20-fold cross-validation as shown in Table 14 and illustrated in Fig. 6. We also indicate a one standard deviation error of the proportions of labels found among the 20 splits in the data.

Using the derived score of the gold standard data set as training and test data, we obtained the following results for the 20-fold cross-validation as shown in Table 15 and illustrated in Fig. 7. The mean absolute error averaged over all cross-validation runs was 0.54, and its standard deviation was 0.04.

7.4 Error analysis

A vocabulary-based annotator such as SentiStrength is easily fooled by its inability to detect negation:

"Mutta ei siellä mitään kamalaa ole!"
But there is nothing horrible!
A pos, B pos, C pos, SentiStrength maximally negative

Alternatively, lacking understanding of compounds, as in this case where it responds to the "horror" in "horror movies":

"Kiistämättä kyllä parhaita kauhuelokuva aikakausia!"
Undeniably one of the best periods for horror movies!
A pos, B pos, C pos, SentiStrength maximally negative

Errors made by neural networks, such as our baseline CNN model, are harder to interpret, except they do not appear to be due to a finite convolution kernel (up to a maximum of five words). However, frequently used smileys and emoticons probably attract a dominant sentiment value when training the neural network classifier.

"Haluan olla se iloinen tyttö pitkästä aikaa. :("
I want to be that happy girl I haven't been for a long time :(
A pos, B pos, C pos, CNN negative

The corpus contains quite a few examples of sarcasm and jest, and one sometimes wonders if the CNN models did not get this more often than the human annotators:

"äänekäs sovinistiörkki! ;)"
You loud chauvinist orc! ;)
A neg, B neg, C neg, CNN positive

8 Discussion and conclusion

In our survey of previous work, we noted that there were only two data sets for sentiment analysis of movie subtitles available for Finnish but no large-scale social media data set with sentiment polarity annotations. This publication remedies this shortcoming by introducing a 27,000-sentence data set annotated independently with sentiment polarity by three native annotators. The same three annotators annotated the whole data set. This is in contrast to other data sets, which have usually been annotated piecemeal by many annotators. Being university students, the annotators belong to a similar demographic, which might introduce some bias. However, bias

detection is a research topic of its own and our resource with consistent annotations by one demographic is a valuable starting point for such research.

Our data set also provides a unique opportunity for further studies of annotator behavior over the sample annotation order, e.g., the human inter-annotator agreement seems to increase without coordination. One can speculate that the annotators became more proficient in their opinion mining towards the end, leading to a convergence in their judgments when getting accustomed to the overall style in the data set. There is also an indication that the annotators at some point began to suffer from annotator fatigue because their rate of neutral judgments increased towards the end, although a verification annotation of a random 1000-sentence subset indicates that there is no real increase of neutral statements towards the end of the data set. This means that it is a question for further studies to determine whether the annotators over-interpreted the sentiments in the beginning or under-interpreted them towards the end.

In addition, we test the data set by providing two baselines validating the usefulness of the data set. Both the baseline validators seem to agree with the verification annotation, indicating that the polarity distribution over the whole data set is rather even, which also indicates that any of the splits provided in the data set will give a rather similarly useful test result because all the splits have data from the whole data set compensating from any potential unevenness in the sentiment judgments of the human annotators over the sample annotation order. The data set is available through the Language Bank of Finland.[12]

Acknowledgements We would like to thank the annotators for their time and FIN-CLARIN and the Language Bank of Finland for access to the data, and the anonymous reviewers for their remarks that helped us clarify the creation process of the data set.

Funding Open Access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

Abdul-Mageed, M., & Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 718–728).

[12] http://urn.fi/urn:nbn:fi:lb-2023012701.

Abdulla, N. A., Al-Ayyoub, M., & Al-Kabi, M. N. (2014). An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence, 1*(1–2), 103–113.

Aller Media Ltd. (2019). The Suomi24 sentences corpus 2001–2017, Korp version 1.1. Retrieved from http://urn.fi/urn:nbn:fi:lb-2020021803.

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics, 10*(11), 1348.

Apidianaki, M., Tannier, X., & Richart, C. (2016). Datasets for aspect-based sentiment analysis in French. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* European Language Resources Association (ELRA), Paris, France.

Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)* (pp. 1–5). IEEE.

Boland, K., Wira-Alam, A., & Messerschmidt, R. (2013). *Creating an annotated corpus for sentiment analysis of german product reviews, GESIS-technical reports* (Vol. 2013/05). GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim.

Bostan LAM., & Klinger R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2104–2119).

City Digital Group. (2021). The Suomi24 sentences corpus 2001–2020, Korp version. Retrieved from http://urn.fi/urn:nbn:fi:lb-2021101525.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., & Wiegand, M. (2012). MLSA—a multi-layered reference corpus for German sentiment analysis. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 3551–3556). European Language Resources Association (ELRA), Istanbul, Turkey. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/125_Paper.pdf.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Einolander, J. (2019). Deeper customer insight from NPS-questionnaires with text mining—comparison of machine, representation and deep learning models in Finnish language sentiment classification. G2 pro gradu, diplomityö, Aalto University. Retrieved from http://urn.fi/URN:NBN:fi:aalto-201904072554.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Feng, Y., & Wan, X. (2019). Learning bilingual sentiment-specific word embeddings without cross-lingual supervision. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota* (pp. 420–429). https://doi.org/10.18653/v1/N19-1040, https://aclanthology.org/N19-1040.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 470–478).

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR), 49*(2), 1–41.

Hämäläinen, M., & Alnajjar, K. (2019). Let's FACE it. Finnish poetry generation with aesthetics and framing. In *Proceedings of the 12th international conference on natural language generation, Association for Computational Linguistics, Tokyo, Japan* (pp. 290–300). https://doi.org/10.18653/v1/W19-8637, https://aclanthology.org/W19-8637.

Hämäläinen, M., & Alnajjar, K. (2021). The current state of finnish nlp. In *Proceedings of the seventh international workshop on computational linguistics of uralic languages* (pp. 65–72)

Harju, A. (2018). Suomi24-keskustelut kohtaamisten ja törmäysten tilana. *Media & viestintä, 41*(1), 51–74.

Hellström, R. (2022). Aspect based sentiment analysis in Finnish. Master's thesis, Aalto University. School of Science. Retrieved from http://urn.fi/URN:NBN:fi:aalto-202202061750.

Jantunen, J. H. (2018). Homot ja heterot suomi24: Ssä: Analyysi digitaalisista diskursseista. *Puhe ja kieli, 38*(1), 3–22.

Jussila, J., Vuori, V., Okkonen, J., & Helander, N. (2017). Reliability and perceived value of sentiment analysis for twitter data. In *Strategic innovative marketing* (pp. 43–48). Springer.

Kajava, K. (2018). Cross-lingual sentiment preservation and transfer learning in binary and multi-class classification. Master's thesis, University of Helsinki.

Kajava, K., Öhman E., Piao H., Tiedemann J. (2020). Emotion preservation in translation: Evaluating datasets for annotation projection. In *DHN* (pp. 38–50)

Karttunen, J. (2021). Predicting omx helsinki stock prices using social media sentiment of finnish retail investors. Master's thesis, Lappeenranta-Lahti University of Technology LUT.

Kaustinen, J. (2018). Sentiment analysis of Finnish movie reviews: Extracting sentiment from texts in a morphologically rich language. Master's thesis, Åbo Akademi.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics, Doha, Qatar* (pp. 1746–1751). Retrieved from https://www.emnlp2014.org/papers/pdf/EMNLP2014181.pdf.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Tech. rep., University of Pennsylvania. Retrieved from http://repository.upenn.edu/asc_papers/43.

Ku, L. W., Lo, Y. S., & Chen, H. H. (2007). Test collection selection and gold standard generation for a multiply-annotated opinion corpus. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 89–92).

Kula, S., Choraś, M., Kozik, R., Ksieniewicz, P., & Woźniak, M. (2020). Sentiment analysis for fake news detection by means of neural networks. In *International conference on computational science* (pp. 653–666). Springer.

Kuutila, M., Mäntyl a, M.V., & Claes, M. (2020). *Chat activity is a better predictor than chat sentiment on software developers productivity, Association for Computing Machinery, New York, NY, USA* (pp. 553–556). https://doi.org/10.1145/3387940.3392224.

Lagus, K., Ruckenstein, M., Pantzar, M., & Ylisiurua, M. (2016). Suomi24: Muodonantoa aineistolle. No. 10 in Valtiotieteellisen tiedekunnan julkaisuja, Helsingin yliopisto, Suomi.

Leuhu, T. (2014). Sentiment analysis using machine learning. Master's thesis, Tampere University of Technology.

Lindén, K., Jauhiainen, T., & Hardwick, S. (2020). FinnSentiment, source. Retrieved from http://urn.fi/urn:nbn:fi:lb-2020111001.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), European Language Resources Association* (pp. 923–929).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on twitter. *Journal of Advertising, 46*(2), 236–247.

Luo, H., Ji, L., Li, T., Jiang, D., & Duan, N. (2020). GRACE: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In *Findings of the association for computational linguistics: EMNLP 2020, Association for Computational Linguistics* (pp. 54–64). https://doi.org/10.18653/v1/2020.findings-emnlp.6, https://aclanthology.org/2020.findings-emnlp.6.

Määttä, S. K., Suomalainen, K., & Tuomarla, U. (2020). Maahanmuuttovastaisen ideologian ja ryhmäidentiteetin rakentuminen suomi24-keskustelussa. Virittäjä.

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.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113.

Mikolov, T., Chen, K., Corrado, G., & Jeffrey, D. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL].

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *SemEval@NAACL-HLT*

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2019). Semeval-2016 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.01973.

Nukarinen, V. (2018). Automated text sentiment analysis for Finnish language using deep learning. Master's thesis, Tampere University of Technology.

Öhman, E. (2020). Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task. In *DHN* (pp. 293–301).

Öhman, E. (2021). Self &; feil: Emotion and intensity lexicons for finnish. https://doi.org/10.48550/ARXIV.2104.13691, https://arxiv.org/abs/2104.13691.

Öhman, E., & Kajava, K. (2018). Sentimentator: Gamifying fine-grained sentiment annotation. In *DHN* (pp. 98–110).

Öhman, E., Honkela, T., & Tiedemann, J. (2016). The challenges of multi-dimensional sentiment analysis across languages. In *Proceedings of the workshop on computational modeling of people's opinions, personality, and emotions in social media (PEOPLES)* (pp 138–142).

Öhman, E., Kajava, K., Tiedemann, J., & Honkela, T. (2018). Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 24–30).

Öhman, E., Pàmies, M., Kajava, K., & Tiedemann, J. (2020). Xed: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th international conference on computational linguistics (COLING 2020)*.

Paavola, J., & Jalonen, H. (2015). An approach to detect and analyze the impact of biased information sources in the social media. In *ECCWS2015-proceedings of the 14th European conference on cyber warfare and security* (p. 213).

Paavola, J., Helo, T., Jalonen, H., Sartonen, M., & Huhtinen, A. (2016a). Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media. *Journal of Information Warfare, 15*(4), 100–111.

Paavola, J., Helo, T., Sartonen, H. J. M., & Huhtinen, A. M. (2016b). The automated detection of trolling bots and cyborgs and the analysis of their impact in the social media. In *ECCWS2016-proceedings of the 15th European conference on cyber warfare and security, Academic Conferences and publishing limited* (p. 237).

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Information Retrieval, 2*(1–2), 1–135.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3–33). Elsevier.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland* (pp. 27–35). https://doi.org/10.3115/v1/S14-2004, https://www.aclweb.org/anthology/S14-2004.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 486–495).

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., & De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *10th international workshop on semantic evaluation (SemEval 2016)*.

Rautiainen, A., & Luoma-aho, V. (2021). Reputation and financial reporting in Finnish public organizations. *Journal of Public Budgeting, Accounting & Financial Management, 33*, 487–511.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop* (pp. 43–48).

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). SemEval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland* (pp. 73–80). https://doi.org/10.3115/v1/S14-2009.

Rosenthal, S., Mohammad, S. M., Nakov, P., Ritter, A., Kiritchenko, S., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. arXiv preprint arXiv:1912.02387.

Rosenthal, S., Farra, N., & Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.00741.

Seki, Y., Evans, D. K., Ku, L. W., Chen, H. H., Kando, N., Lin, C. Y. (2007). Overview of opinion analysis pilot task at ntcir-6. In *NTCIR*

Seki, Y., Evans, D. K., Ku, L. W., Sun, L., Chen, H. H., Kando, N., Lin, C. Y. (2008). Overview of multilingual opinion analysis task at ntcir-7. In *NTCIR*.

Seki, Y., Ku, L. W., Sun, L., Chen, H. H., & Kando, N. (2010). Overview of multilingual opinion analysis task at ntcir-8. In *Proc. of the Seventh NTCIR Workshop*.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63*(1), 163–173.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. *Lrec, 2012*, 2214–2218.

Vankka, J., Myllykoski, H., Peltonen, T., & Riippa, K. (2019). Sentiment analysis of finnish customer reviews. In *2019 sixth international conference on social networks analysis* (pp. 344–350). IEEE: Management and Security (SNAMS).

Vankka, J., Vesselkov, A., Myllykoski, H., & Kosomaa, O. (2021). Framework for analyzing and visualizing topics and sentiments on social media: the case of mh 17 tweets. In *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)* (pp. 257–266). https://doi.org/10.1109/ICBDA51983.2021.9403069.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. https://doi.org/10.48550/ARXIV.1912.07076, https://arxiv.org/abs/1912.07076.

y Montse Cuadros y Seán Gaines y German Rigau, R. A. (2013). Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural, 51*, 215–218.

Ylilauta. (2015). Ylilauta Corpus. Retrieved from http://urn.fi/urn:nbn:fi:lb-2015031802.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.