# Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian

**Petter Mæhlum, Jeremy Barnes, Lilja Øvrelid, and Erik Velldal**
University of Oslo
Department of Informatics
{pettemae,jeremycb,liljao,erikve}@ifi.uio.no

## Abstract

This paper documents the creation of a large-scale dataset of evaluative sentences – i.e. both subjective and objective sentences that are found to be sentiment-bearing – based on mixed-domain professional reviews from various news-sources. We present both the annotation scheme and first results for classification experiments. The effort represents a step toward creating a Norwegian dataset for fine-grained sentiment analysis.

## 1 Introduction

Sentiment analysis is often approached by first locating the relevant, sentiment-bearing sentences. Traditionally, one has distinguished between subjective and objective sentences, where only the former were linked to sentiment (Wilson, 2008). Objective sentences typically present facts about the world, whereas subjective sentences express personal feelings, views, or beliefs. More recently, however, it has become widely recognized in the literature that subjectivity should not be equated with opinion (Liu, 2015): On the one hand, there are many subjective sentences that do not express sentiment, e.g., *I think that he went home*, and on the other hand there are many objective sentences that do, e.g., *The earphone broke in two days*, to quote some examples from Liu (2015). Additionally, sentences often contain several polarities in a single sentence, which complicates the labeling of a full sentence as positive or negative.

This paper documents both the annotation effort and first experimental results for sentence-level evaluative labels added to a subset of the data in the Norwegian Review Corpus (NoReC) (Velldal et al., 2018), a corpus of full-text reviews from a range of different domains, collected from several of the major Norwegian news sources. The annotated subset, dubbed NoReC$_{eval}$, covers roughly 8000 sentences across 300 reviews and 10 different thematic categories (literature, products, restaurants, etc.).

Sentences are labeled to indicate whether they are *evaluative*, i.e. where they are intended by the author (or some other opinion holder) to serve as an evaluation or judgment. They are not, however, annotated with respect to positive/negative polarity. The reason for this is that polarity is often mixed at the sentence-level. Hence, we defer annotating polarity to a later round of phrase-level annotation. Although most of the sentences labeled as evaluative will be subjective and personal, they can also include objective sentences. Moreover, our annotation scheme singles out a particular category of evaluative sentences called *fact-implied non-personal*, following the terminology of Liu (2015). Evaluative sentences are also further sub-categorized as to whether they are considered *on-topic* with respect to the object being reviewed, and whether they express the *first-person* view of the author.

The annotation scheme is described in further detail in Sections 3 and 4. We start, however, by briefly outlining relevant previous work and background in Section 2. In Section 5 we describe more practical aspects of the annotation procedure and go on to analyze inter-annotator agreement in Section 6, before Section 7 summarizes the resulting dataset. In Section 8, we analyze the corpus experimentally and present a series of preliminary classification experiments using a wide range of state-of-the-art sentiment models including CNNs, BiLSTMs and self-attention networks, before we in Section 9 conclude and outline some remaining avenues for future work. The dataset and the annotation guidelines are made available, along with code for replicating the experiments.[1]

[1] https://github.com/ltgoslo/norec_eval



*Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 121–130
Turku, Finland, 30 September – 2 October, 

## 2 Background and related work

In this section we briefly review some of the previous annotation efforts (for English) that are most relevant for our work.

Toprak et al. (2010) present a sentiment-annotated corpus of consumer reviews. In a first pass, sentences are annotated with respect to relevancy to the overall topic and whether they express an evaluation. In a second pass, sentences that were marked as relevant and evaluative are further annotated with respect to whether they are *opinionated* (i.e. express a subjective opinion) or *polar-facts* (i.e. factual information that implies evaluation). In addition to evaluations, they also identify sources (opinion holders), targets (the entity or aspect that the sentiment is directed towards), modifiers, positive/negative polarity and strength, and anaphoric expressions.

Also working with review data, Scheible and Schütze (2013) present a simplified annotation scheme which appears similar in spirit to the first pass of annotation described by Toprak et al. (2010). Scheible and Schütze (2013) annotate sentences with respect to what they call *sentiment relevance*, indicating whether they are informative for determining the sentiment of a document. Sentiment relevant sentences can be either subjective or objective, but must be on topic and convey some evaluation of the object under review.

Van de Kauter et al. (2015) present a fine-grained scheme for annotation of polar expressions at the sub-sentential level. They distinguish between two types of sentiment; *explicit* sentiment on the one hand, corresponding to private states, and *implicit* sentiment on the other, corresponding to factual information that implies a positive/negative evaluation (van de Kauter et al., 2015). The latter category corresponds to what is refered to as *polar-facts* by Toprak et al. (2010) or *objective polar utterances* by Wilson (2008). The annotations of van de Kauter et al. (2015) also identify sources, targets, and modifiers. Acknowledging that the distinction between implicit/explicit sentiment is not always clear cut, polar expressions are labeled with a graded numerical value indicating a continuum ranging from objective to subjective.

Liu (2015) proposes various sub-categorizations of what he calls *opinionated* expressions along several dimensions. Among the most relevant for our work is the distinction between *subjective* and *fact-implied opinions*. The subjective expressions are further sub-categorized as either *emotional* or *rational*, and the fact-implied can be either *personal* or *non-personal* (Liu, 2015). In the order they are listed above, these sub-categorizations can perhaps be seen to correspond to four bins of the subjective–objective continuum defined by van de Kauter et al. (2015). Liu (2015) also differentiates between first-person and non-first-person opinions, where non-first-person indicates that the opinion is held by someone other than the author of the sentence.

In the next section we describe the choice of label categories used in our sentence-level annotation of NoReC reviews.

## 3 Annotation scheme

Our annotation approach corresponds to some degree to that of Scheible and Schütze (2013) or the first step described by Toprak et al. (2010) – see discussion above – in that we assign labels only at the sentence-level and without marking polarity (as this might be mixed at the sentence-level), and include both subjective and objective sentences. However, our approach is slightly more fine-grained in that we also explicitly annotate evaluative sentences with respect to being on-topic or not, and with respect to expressing a first-person opinion of the author or not. Finally, we also single out one particular sub-class of evaluative sentences, namely those that in the terminology of Liu (2015) are fact-implied non-personal. These sentences might require special treatment, where proper identification might be more dependent on taking the overall domain and discourse context into account (Liu, 2015). In this section we provide more details and examples for the various label types in our annotation scheme.

**Evaluative** Following Toprak et al. (2010), we use the term *evaluative* to refer to any sentence that expresses or implies a positive or negative evaluation, regardless of its subjectivity. An example of an evaluative sentence can be found in (1) below which contains the positive evaluation signaled by the adjective *lekkert* 'tastefully'.

(1) *Det hele var også lekkert presentert.*
The whole was also tastefully presented.
'Everything was tastefully presented.'

Our `EVAL` label roughly comprises the three opinion categories described by Liu (2015) as

emotional, rational and fact-implied personal. Sentences including emotional responses (arousal) are very often evaluative and involve emotion terms like e.g. *elske* 'love', *like* 'like' or *hate* 'hate'. Sentences that lack the arousal we find in emotional sentences may also be evaluative, for instance by indicating worth and utilitarian value, e.g. *nyttig* 'useful' or *verdt (penger, tid)* 'worth (money, time)'.

**Evaluative fact-implied non-personal** There are actually two types of evaluative sentences in our scheme: simply *evaluative* (labeled EVAL) as in (1) above, or the special case of *evaluative fact-implied non-personal* (FACT−NP).

A sentence is labeled as FACT−NP when it is a fact or a descriptive sentence but evaluation is implied, and the sentence does not involve any personal experiences or judgments. (In contrast, objective sentences expressing personal experiences – so-called *fact-implied personal* in the terminology of Liu (2015) – are not seen as objective to the same degree, and are labeled as EVAL.) FACT−NP-labeled sentences are usually understood to be evaluative because we interpret them based on common (societal, cultural) background knowledge, and they are often highly context dependent. The example in (2) illustrates a FACT−NP-labeled sentence which simply states factual information, however, within the context of a car review, it clearly expresses a positive evaluation.

(2) *178 hestekrefter.*
178 horsepowers.
'178 horsepower.'

Note that the definition of FACT−NP departs from what at first might appear like similar categories reported in the literature, like factual *implicit* sentiment (van de Kauter et al., 2015), *polar-facts* (Toprak et al., 2010) or *objective polar utterances* (Wilson, 2008), in that it does not include so-called *personal* fact-implied evaluations (Liu, 2015). This latter class is in our scheme subsumed by EVAL. The reason for this is that we found them to have a more explicit and personal nature, separating them from the purely objective FACT−NP sentences described above.

**Non-evaluative** Sentences that do not fall into either of these two categories (EVAL and FACT−NP) are labeled non-evaluative (NONE). An example of this category can be found in (3), which is taken from a restaurant review. Even though this sentence clearly describes a personal experience, it is still a factual statement that does not express any sort of evaluation.

(3) *Jeg har aldri spist den oransje*
I have never eaten the orange
*varianten av sorten, sa Fredag.*
variant of kind.the, said Fredag.
'I have never tasted the orange kind, said Fredag'

**On-topic or not** Sentences that are identified as evaluative, in either the EVAL or FACT−NP sense, are furthermore labeled with respect to two other properties: (*i*) whether the author is the one expressing the evaluation, and (*ii*) whether the evaluation is on topic or not.

Sentences that are not-on-topic are labeled ¬OT. For an example, see (4), where the review is about a music album, but the sentence expresses an evaluation about the author upon whose book the album is based, and does not reflect the reviewer's evaluation of the album itself.

(4) *Jeg liker Aune Sand.*
I like Aune Sand
'I like Aune Sand [name of author].'

The class of sentiment-bearing sentences that are not considered relevant or on-topic are typically not marked in other annotation efforts, e.g. by Toprak et al. (2010) or Scheible and Schütze (2013). However, from a modeling perspective, we expect it will be difficult in practice to correctly identify evaluative sentences that are on-topic while leaving out those that are not, at least without going beyond the standard sentence-level models typically applied in the field today and move towards more discourse-oriented modeling. By explicitly labeling the not-on-topic cases we are able to quantify this effect, both with respect to human annotations and system predictions.

**First person or not** Sentences where the author is not the holder of the evaluation, are labeled ¬FP ('not-first-person'). An example is provided in (5) where the holder of the opinion is not the author of the review, but rather the subject noun phrase *ekte astronauter* 'real astronauts'.

(5) *Ekte astronauter har også sett*
real astronauts have also seen
*filmen og skryter hemningsløst av*
movie.the and boast unrestrainedly of
*dens autentisitet*
its authenticity

'Real astronauts have also seen the movie and boast highly of its authenticity'

**Mixed class sentences** A sentence may include several types of evaluative expressions. In these cases, we label a sentence as EVAL if it contains both EVAL and FACT-NP, as in example (6) below.

(6) *Dette gir et gjennomsnitt på 27,3*
this gives an average on 27,3
*MB/sek som er meget bra.*
MB/sec which is very good
'This gives us an average of 27,3 MB / sec, which is very good.'

Similarly, we refrain from labeling ¬OT and ¬FP if a sentence contains any sentiment expression that is first-person or on topic respectively.

## 4 Annotation challenges / special cases

Below, we provide some more details about particular annotation decisions related to various special cases, including some challenges.

**Modality** In our annotation guidelines, the treatment of modals depends on the specific modal verb in use. In particular, we found that some modals like *burde* 'should' are frequently used to indicate evaluation, as in the example (7) below.

(7) *Hun burde hatt med seg en*
She should had with herself an
*opplevelse i tillegg.*
experience in addition.
'On top of this she should have brought with her an experience.'

**Conditionals** Conditional sentences also require special attention. In particular, so-called irrealis sentences, i.e., sentences that indicate hypothetical situations, have been excluded in some previous sentence-level annotation efforts (Toprak et al., 2010), but we wish to include them as long as they clearly indicate evaluation. A seemingly common use of irrealis is to indicate negative evaluation by expressing a future condition, indicating that the current situation is less optimal, as in (8) below.

(8) *Bare Elvebredden får nok arbeidskraft*
Only Elvebredden gets enough work-power
*[...] gleder Robinson & Fredag*
[...] look-forward Robinson & Fredag
*seg til å komme tilbake*
themselves to INF come back
'If only Elvebredden had more waiters, Robinson & Fredag would gladly return'

**Questions** Questions often have a similar role in expressing evaluations as the conditionals discussed above. Often a sentence may question some aspect of the object in question, also indicating a negative evaluation of the current state of the object, as in (9) below, labeled EVAL.

(9) *Et "mimrespill" skal vel stimulere*
A memory-game should well stimulate
*mer enn korttidsminnet?*
more than shortterm.memory.the?
'Shouldn't a "memory game" stimulate more than the short term memory?'

**Cross-sentential evaluation** An evaluative expression may sometimes span across several sentences. Since our annotation is performed at the sentence-level, annotations may not span across sentences. We decided to label adjacent sentences that were strongly related identically. In examples (10) and (11) below, for instance, the first sentence contains a general comment about the action scenes penned by a given book author, but this is tied to the topic of the review (the author's new book *Gjenferd* 'Ghost') only in the sentence following it. In our annotation, these two sentences were both annotated as EVAL.

(10) *Min største innvending er at*
my biggest objection is that
*actionscenene til Nesbø har en*
action.scenes.the of Nesbø has a
*tendens til å få noe*
tendency to INF get something
*tegneserieaktig overdrevent over seg.*
cartoon.like exaggerated over themselves
'My biggest objection is that Nesbø's action scenes have a tendency to give an exaggerated cartoon-like expression.'

(11) *Det gjelder også i "Gjenferd".*
That applies also in "Gjenferd"
'That also applies in "Gjenferd" [book title].'

Other examples of evaluative expressions spanning sentences are lists of reasons following or preceding a more clearly evaluative expression, and sentences where the target and polar expression are split, as in a question–answer structure.

**External objective evaluation** Another challenging type of sentence encountered during annotation are sentences where the author refers to prizes or evaluations by people other than the author, as in (12) below. These expressions are marked as ¬FP, but evaluation-wise they can be

seen from two angles: Is the author using the phrase to express an explicit positive evaluation, in which case it would be marked as EVAL, or is the author reporting a fact, in which case it is marked as FACT-NP. The same problem applies to words like *populær* 'popular' or *folkekjær* 'loved by the people', although these words tend towards EVAL, while nominations like in (12) tend towards FACT-NP.

(12) *[...] er både Ejiofor og Fassbender*
[...] are both Ejiofor and Fassbender
*Oscar-nominert.*
Oscar-nominated .
'[...]] both Ejiofor and Fassbender have been Oscar-nominated.'

In this case, the evaluation has been performed by a different group of people at an earlier stage and the evaluation is also not of the object being reviewed, and is therefore marked as ¬OT, ¬FP and FACT-NP.

**Higher-level topic evaluation** At times the annotators also found sentences where the evaluation is at a higher ontological level than the object being reviewed, as in sentence (13), where the review is about a specific edition of a series of games called *Buzz*, but the evaluation is about the series as a whole.

(13) *Da tror jeg Buzz kan fenge i*
Then think I Buzz can captivate in
*mange år til [...].*
many years more [...]
'Then I think Buzz [game] can captivate for many more years'

In these cases, it was decided that as long as the object being reviewed is a close subclass of the target of the evaluation, it is reasonable to assume that the author wrote this sentence in order to say something about the overall quality of the actual object under review, and thus the sentence above is labeled EVAL.

## 5 Annotation procedure

Annotation was performed using the WebAnno tool (Eckart de Castilho et al., 2016), and annotators were able to see the whole review in order to judge sentences in context. There were five annotators in total (students with background in linguistics and language technology) and all sentences were doubly-annotated. In cases of disagreement, another of the annotators would consider the sentence a second time and resolve the conflict. Problematic sentences would be discussed at a meeting with all annotators present.

The annotation guidelines were fine-tuned in three rounds using two sets of texts. The first set contained 10 texts, representing each of the thematic categories in NoReC, in order to provide the annotators with as much variation as possible. These texts were annotated by two of the annotators, and the results were discussed, forming the basis of the guidelines. The same annotators then annotated a second set of 8 texts, trying to strictly adhere to the guidelines. After a second fine-tuning, the remaining annotators would annotate the first set, and the guidelines were again fine-tuned in accordance with the new disagreements. These texts are not included when calculating the agreement scores reported below.

## 6 Inter-annotator agreement

Inter-annotator agreement scores for the main three categories EVAL, FACT-NP, and NONE are presented in Table 1, calculated as $F_1$-scores between pairs of annotators on the complete set of sentences. We find that agreement among the annotators is high for the EVAL sentences and for the overall score. Agreement is much lower for the FACT-NP label, however, likely reflecting the fact that these sentences have no clear sentiment expression, with interpretation more heavily depending on context and domain-specific knowledge.

We also computed annotator agreement for the attribute categories ¬OT and ¬FP, restricted to the subset of sentences labeled EVAL,[2] yielding $F_1$ of 0.59 and 0.56, respectively. In other words, we see that the agreement is somewhat lower for these subcategories compared to the top-level label EVAL. Possible reasons for this might be that although problems with these attributes seem to be resolved quickly in annotator meetings, they might pose difficulties to the individual annotator, as sometimes these attributes can be context dependent to an extent that makes them difficult to infer from the review text by itself.

Kenyon-Dean et al. (2018) problematizes a practice often seen in relation to sentiment annotation, namely that complicated cases – e.g. sentences were there is annotator disagreement – are discarded from the final dataset. This makes the

[2]For the FACT-NP subset there were too few instances of these attributes (prior to adjudication) for agreement to be meaningfully quantified; 1 for ¬OT and 0 for ¬FP.

| EVAL | FACT-NP | NONE | all |
|---|---|---|---|
| 0.84 | 0.22 | 0.87 | 0.82 |

Table 1: $F_1$ inter-annotator agreement for each top-level label.

data non-representative of real text and will artificially inflate classification results on the annotations. In our dataset, we not only include the problematic cases, but also explicitly flag sentences for which there was disagreement among annotators (while also indicating the resolved label). This can be of potential use for both error analysis and model training, as we will also see in Section 8.3. Finally, note that we also found interesting differences in agreement across review domains and this too is something we return to when discussing experimental results in Section 8.3.

## 7 Corpus statistics

Table 2 presents the distribution of the annotated classes (EVAL, FACT-NP and NONE), as well as the attributes ¬OT and ¬FP in terms of absolute number and proportion of sentences across the different review domains (screen, music, literature, etc.). The resulting corpus contains a total of 298 documents and 7961 total sentences.

In general, we may note that there is a large proportion of evaluative sentences in the corpus, a fact which is unsurprising given the review genre. EVAL sentences are in a slight majority in the corpus (just above 50%) followed by NONE which accounts for 46% of the sentences, while the FACT-NP label makes up a little less than 4% of the sentences.

We observe that the evaluative sentences (EVAL or FACT-NP) are not evenly distributed across the different thematic categories. The category with the highest percentage of evaluative sentences – restaurants – tend to be written in a personal style, with vivid descriptions of food and ambience. In contrast, stage reviews tend to be written in a non-personal style, largely avoiding strong evaluations. Unsurprisingly, the product category has a higher number of FACT-NP sentences, as they contain several objective but evaluative product descriptions. The low proportion of EVAL sentences found in the literature category is somewhat surprising, as one would not normally consider literature reviews as especially impersonal. However, music reviews in this corpus tend to be written in a personal, informal style, which is reflected in the high rate of EVAL sentences.

The corpus contains a total of 396 ¬OT sentences and 109 ¬FP sentences. Most of the evaluative sentences are thus on topic, and most evaluations belong to the author. The percentages of the attributes ¬OT and ¬FP are quite evenly distributed among the different domains, with the exception of one apparent outlier: the 31.33% of ¬FP sentences in the sports domain. This is probably due to the interview-like style in one of the reviews, reporting the evaluations of several different people. Reviews about video games seem to have a slightly higher percentage of ¬OT sentences. This could be due to a large number of comparisons with earlier games and different gaming consoles in these texts.

## 8 Experiments

In this section we apply a range of different architectures to provide first baseline results for predicting the various labels in the new corpus. Data splits for training, validation and testing are inherited from NoReC.

### 8.1 Models

We provide a brief description of the various classifiers below. Additionally, we provide a majority baseline which always predicts the EVAL class as a lower bound. Note that all classifiers except the bag-of-words model take as input 100 dimensional fastText skipgram embeddings (Bojanowski et al., 2016), trained on the NoWaC corpus (Guevara, 2010), which contains over 680 Million tokens in Bokmål Norwegian. The pre-trained word embeddings were re-used from the NLPL vector repository[3] (Fares et al., 2017).

**BOW** learns to classify the sentences with a linear separation estimated based on log likelihood optimization with an L2 prior using a bag-of-words representation.

**AVE** (Barnes et al., 2017) uses the same L2 logistic regression classifier as BOW, but instead using as input the average of the word vectors from a sentence.

**CNN** (Kim, 2014) is a single-layer convolutional neural network with one convolutional layer

[3] http://vectors.nlpl.eu/repository/

| Domain | Docs | Sents | EVAL # | EVAL % | FACT-NP # | FACT-NP % | NONE # | NONE % | ¬OT # | ¬OT % | ¬FP # | ¬FP % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Screen | 110 | 2895 | 1359 | 46.94 | 50 | 1.73 | 1486 | 51.33 | 160 | 11.36 | 20 | 1.42 |
| Music | 101 | 1743 | 1055 | 60.53 | 48 | 2.75 | 640 | 36.72 | 100 | 9.07 | 23 | 2.09 |
| Literature | 35 | 930 | 327 | 35.16 | 31 | 3.33 | 572 | 61.51 | 50 | 13.97 | 18 | 5.03 |
| Products | 22 | 1156 | 619 | 53.55 | 127 | 10.99 | 410 | 35.47 | 36 | 4.83 | 10 | 1.34 |
| Games | 13 | 520 | 278 | 53.46 | 23 | 4.42 | 219 | 42.12 | 37 | 12.29 | 6 | 1.99 |
| Restaurants | 6 | 268 | 167 | 62.31 | 10 | 3.73 | 91 | 33.96 | 4 | 2.26 | 6 | 3.39 |
| Stage | 8 | 264 | 100 | 37.88 | 6 | 2.27 | 158 | 59.85 | 7 | 6.60 | 0 | 0.0 |
| Sports | 2 | 149 | 78 | 52.35 | 5 | 3.36 | 66 | 44.3 | 2 | 2.41 | 26 | 31.33 |
| Misc | 1 | 36 | 20 | 55.56 | 0 | 0.0 | 16 | 44.44 | 0 | 0.0 | 0 | 0.0 |
| Total | 298 | 7961 | 4003 | 50.28 | 300 | 3.77 | 3658 | 45.95 | 396 | 9.20 | 109 | 2.53 |

Table 2: Distribution of documents, sentences and labels across the thematic categories of reviews. Note that the percentages for ¬OT and ¬FP are relative to evaluative (EVAL or FACT-NP) sentences.

on top of pre-trained embeddings. The embedding layer in convoluted with filters of size 2, 3, and 4 with 50 filters for each size and then 2-max pooled. This representation is then passed to a fully connected layer with *ReLU* activations and finally to a softmax layer. Dropout is used after the max pooling layer and *ReLU* layer for regularization.

**BILSTM** is a one-layer bidirectional Long Short-Term Network (Graves et al., 2005) with word embeddings as input. The contextualized representation of each sentence is the concatenation of the final hidden states from the left-to-right and right-to-left LSTM. This representation is then passed to a softmax layer for classification. Dropout is used before the LSTM layers and softmax layers for regularization.

**SAN** is a one-layer self-attention network (Vaswani et al., 2017) with relative position representations (Shaw et al., 2018) and a single set of attention heads, which was previously shown to perform well for sentiment analysis (Ambartsoumian and Popowich, 2018). The network uses a variant of the attention mechanism (Bahdanau et al., 2014) which creates contextualized representations of the original input sequence, such that the contextualized representations encode both information about the original input, as well as how it relates to all other positions.

### 8.2 Experimental Setup

We apply the models to five experimental setups. The main task is to classify each sentence as *evaluative* (EVAL), *fact-implied non-personal* (FACT-NP), or *non-evaluative* (NONE). In order to provide a view of how difficult it is to model the secondary properties mentioned in Section 3, two additional binary classification tasks are performed; determining if the sentence is on topic (OT) and if the opinion expressed is from a first-person perspective (FP). Only the best performing model from the main experiment above is applied for these subtask, and the model is trained and tested separately on the two subsets of sentences annotated as EVAL and FACT-NP, leading to four binary classification experiments in total.

For all models, we choose the optimal hyper-parameters by performing a random search on the development data. Given that neural models are sensitive to random initialization parameters, we run each neural experiment five times with different random seeds and report means for both per-class and micro $F_1$ in addition to their standard deviation.

| Model | EVAL | FACT-NP | NONE | Overall |
|---|---|---|---|---|
| majority | 66.2 | 0.0 | 0.0 | 49.5 |
| BOW | 69.6 | 0.0 | 64.4 | 65.8 |
| AVE | 75.4 | 0.0 | 70.4 | 71.6 |
| CNN | **76.3** (0.7) | 0.0 (0.0) | 72.2 (0.7) | 73.1 (0.3) |
| BILSTM | 76.1 (0.1) | 6.0 (4.8) | 72.1 (0.1) | 72.7 (0.1) |
| SAN | 76.2 (0.1) | **7.1** (3.1) | **72.3** (0.3) | **73.7** (0.1) |

Table 3: Per class $F_1$ score and overall micro $F_1$ of baseline models on the main classification task. For the neural models mean micro $F_1$ and standard deviation across five runs are shown.

### 8.3 Results

Table 3 shows the results for all models on the main three-way classification task. All classifiers perform better than the majority baseline (at 49.5

$F_1$ overall). Of the two logistic regression classifiers, the AVE model based on averaged embeddings as input performs much better than the standard discrete bag-of-words variant (65.8 vs. 71.6 overall). While the AVE model proves to be a strong baseline, the three neural models have the strongest performance. The CNN achieves the best results on the EVAL class (76.3) and improves 1.8 ppt over AVE on NONE. While overall results are quite even, the strongest model is SAN – the self-attention network – which achieves an overall $F_1$ of 73.7. This model also proves more stable in the sense of having slightly lower variance across the multiple runs, at least compared to the CNN.

The easiest class to predict is EVAL, followed closely by NONE. The most striking result is that is appears very difficult for all models to identify the FACT-NP class. This is largely due to the few examples available for FACT-NP, as well as the fact that FACT-NP sentences do not contain clear lexical features that separate them from EVAL and NONE. This confirms the intuitions presented in Section 3. Only BILSTM and SAN manage to make positive predictions for FACT-NP, but the scores are still very low (with 7.1 $F_1$ being the best) and we see that the variance across runs is high. An analysis of the strongest model (SAN) shows that the model tends to confuse FACT-NP nearly equally with EVAL (15 errors) and NONE (20 errors), while only correctly predicting this category 6 times, suggesting this category is difficult for the models to capture.

**Performance per domain** Table 4 breaks down the $F_1$ score of the SAN model across the different review domains. We observe that there are fairly large differences in performance, and furthermore that these can not simply be explained just by differences in the number of training examples for each domain (cf. the class distributions in Table 2). We see that sentences from the literature reviews appear difficult to classify, despite being relatively well represented in terms of training examples, while the opposite effect can be seen for the games category. The lowest performance is seen for the product reviews, which is unsurprising given that – despite having a high number of examples – it is arguably the most heterogeneous category in the dataset, in addition to having a relatively high proportion of the difficult FACT-NP sentences.

| **Domain** | $F_1$ |
|---|---|
| Screen | 77.5 (2.2) |
| Music | 76.1 (1.3) |
| Literature | 66.0 (1.3) |
| Products | 65.0 (0.8) |
| Games | 77.6 (2.2) |
| Restaurants | 69.6 (1.5) |
| Stage | 70.0 (2.2) |

Table 4: Per domain micro $F_1$ score of the SAN model. Note that the test set does not contain sentences from the Sports or Misc domains.

**Human agreement vs model performance** We also computed the inter-annotator agreement scores per domain, again as pairwise micro $F_1$, and found that while the agreement tends to vary less than model performance, the two scores yield a similar relative ranking of domains in terms of difficulty. For example, the two domains with the highest prediction scores, Games and Screen (with $F_1$ of 77.6 and 77.5, respectively), also have the highest inter-annotator agreement (82.6 and 83.8). The two domains with lowest prediction $F_1$, Products and Restaurants (65.0 and 69.6, respectively), also have the lowest agreement (77.54 and 78.5).

As described in Section 3, while annotator disagreements have been resolved, we have chosen to mark them in the final dataset. An error analysis of the classifier predictions show there is a strong correlation between inter-annotator agreement and errors that the classification models make (using a $\chi^2$ test, $p \ll 0.01$). This suggests that these examples are inherently more difficult, and lead to disagreement for both human and machine learning classifiers.

**On-topic and first-person** Table 5 shows the results of applying the SAN architecture to the four binary tasks. The sentences which are on-topic (OT) and first-person (FP) are the easiest to classify ($F_1$ ranging from 92.8 to 99.4), while the not-on-topic (¬OT) and not-first-person (¬FP) are very difficult (0.0 – 11.3 $F_1$). None of the models are able to correctly predict the ¬FP class. In order to distinguish this class, some kind of co-reference resolution likely needs to be included in the model, as simple lexical information cannot distinguish them from FP. Note, however, that the prediction scores for ¬FP need to be taken with a

| Model | Subset | OT | ¬OT | Avg. | FP | ¬FP | Avg. |
|---|---|---|---|---|---|---|---|
| SAN | EVAL | 93.5 (0.1) | 11.3 (4.3) | 88.5 (1.0) | 99.4 (0.0) | 0.0 (0.0) | 98.9 (0.0) |
| | FACT-NP | 97.2 8 (0.0) | 0.0 (0.0) | 94.6 (0.0) | 92.8 (0.0) | 0.0 (0.0) | 86.5 (0.0) |

Table 5: Per-class and micro $F_1$ for the self-attention network trained to predict whether an example is on topic (OT) or not (¬OT) or whether the opinion is expressed by the first person (FP) or not (¬FP). The models are trained and tested on the subset of sentences annotated as evaluative (EVAL) and fact-implied (FACT-NP).

grain of salt as there are too few instances in the test data to give reliable estimates; 5 in each of the EVAL and FACT-NP subsets. The same is true of the ¬OT predictions for FACT-NP (8 test instances). We see that the network is able to predict to some degree (11.3) the ¬OT class for EVAL, but the absolute score is still low, which also reflects the inter-annotator scores. Once information about aspect or target expressions is added to the data in future annotation efforts, we hope that this might be leveraged to more accurately predict 'ontopicness'.

## 9 Summary and outlook

This paper has described an annotation effort focusing on evaluative sentences in a subset of the mixed-domain Norwegian Review Corpus, dubbed NoReC$_{eval}$. Both subjective and objective sentences can be labeled as evaluative in our annotation scheme. One particular category of objective sentences, conveying so-called *fact-implied non-personal* sentiment, is given a distinct label, as this category might need special treatment when modeling. Evaluative sentences are also assigned labels that indicate whether they are on topic and express a first-person point of view.

The paper also reports experimental results for predicting the annotations, testing a suite of different linear and neural architectures. While the neural models reach a micro $F_1$ of nearly 74 on the three-way task, none of them are able to successfully predict the underrepresented minority-class FACT-NP, misclassifying it nearly equally as often with EVAL as with NONE. Additional experiments show that it is difficult to classify sentences as not-on-topic (¬OT) and not-first-person (¬FP), indicating that important of this in future research on sentiment analysis. Moreover, our error analysis also showed that the cases where annotators disagree (flagged in the data) are also difficult for the classifiers to predict correctly.

Note that, in our annotation scheme, we only annotate sentences as sentiment-bearing (i.e. evaluative), not with positive/negative polarity values, as labeling polarity on the sentence-level only makes sense for sentences that do not contain mixed sentiment. Although such datasets are not uncommon, we argue that this is a rather idealized classification task not in line with the goal of the current effort. In immediate follow-up work, however, we will perform fine-grained sentiment annotation where we label in-sentence sentiment expressions and their polarity, in addition to sources (holders) and targets (aspect expressions). In later iterations we plan to also analyze additional information that can be compositionally relevant to polarity like negation, intensifiers, verbal valence shifters, etc. The dataset and the annotation guidelines are made available, along with code for replicating the experiments.[4]

## Acknowledgements

This work has been carried out as part of the SANT project (Sentiment Analysis for Norwegian Text), funded by the Research Council of Norway (grant number 270908). We also want to express our gratitude to the annotators, who in addition to the first author includes Anders Næss Evensen, Carina Thanh-Tam Truong, Tita Enstad, and Trulz Enstad. Finally, we thank the anonymous reviewers for their valuable feedback.

## References

Artaches Ambartsoumian and Fred Popowich. 2018. Self-attention: A better building block for sentiment analysis neural network classifiers. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 130–139, Brussels, Belgium.

[4] https://github.com/ltgoslo/norec_eval

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12, Copenhagen, Denmark.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, LT4DH@COLING*, pages 76–84, Osaka, Japan.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 271–276, Gothenburg, Sweden.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In *Artificial Neural Networks: Biological Inspirations - ICANN 2005, LNCS 3697*, pages 799–804. Springer-Verlag Berlin Heidelberg.

Emiliano Raul Guevara. 2010. NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, NAACL-HLT, Los Angeles.

Marjam van de Kauter, Bart Desmet, and Véronique Hoste. 2015. The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49:685–720.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies*, pages 1886–1895, New Orleans, Louisiana.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar.

Bing Liu. 2015. *Sentiment analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge, United Kingdom.

Christian Scheible and Hinrich Schütze. 2013. Sentiment relevance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 954–963, Sofia, Bulgaria.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 130–139, Uppsala, Sweden.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Erik Velldal, Lilja Øvrelid, Cathrine Stadsnes Eivind Alexander Bergem, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Theresa Wilson. 2008. Annotating subjective content in meetings. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, pages 2738–2745, Marrakech, Morocco.