A Fine-Grained Sentiment Dataset for Norwegian

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

Abstract

We introduce NoReC$_{fine}$, a dataset for fine-grained sentiment analysis in Norwegian, annotated with respect to polar expressions, targets and holders of opinion. The underlying texts are taken from a corpus of professionally authored reviews from multiple news-sources and across a wide variety of domains, including literature, games, music, products, movies and more. We here present a detailed description of this annotation effort. We provide an overview of the developed annotation guidelines, illustrated with examples, and present an analysis of inter-annotator agreement. We also report the first experimental results on the dataset, intended as a preliminary benchmark for further experiments.

Keywords: Sentiment analysis, opinion mining, Norwegian

1. Introduction

In this work, we describe the annotation of a fine-grained sentiment dataset for Norwegian, analysing opinions in terms of their polar expressions, targets, and holders. The dataset, including the annotation guidelines, is made publicly available[1] and is the first of its kind for Norwegian. The underlying texts are taken from the Norwegian Review Corpus (NoReC) (Velldal et al., 2018) – a corpus of professionally authored reviews across a wide variety of domains, including literature, video games, music, various product categories, movies, TV-series, restaurants, etc. In Mæhlum et al. (2019), a subset of the documents, dubbed NoReC$_{eval}$, were annotated at the sentence-level, indicating whether a sentence contains an *evaluation* or not. These prior annotations strictly indicated evaluativeness and did not include negative or positive polarity, as this can be mixed at the sentence-level.

In the current work, the previous annotation effort has been considerably extended to include the span of polar expressions and the corresponding *targets* and *holders* of the opinion. We also indicate the *intensity* of the *positive or negative polarity* on a three-point scale, along with a number of other attributes of the expressions. The resulting dataset, dubbed NoReC$_{fine}$, comprises almost 8000 sentences (of which roughly half are evaluative) across roughly 300 reviews and includes both subjective and fact-implied evaluations. In addition to discussing annotation principles and examples, we also present the first experimental results. The paper is structured as follows. Section 2 reviews related work, both in terms of related resources for other languages and work on computational modeling of fine-grained opinions. We then go on to discuss our annotation effort in Section 3, where we describe annotation principles, discuss a number of examples and finally present statistics on inter-annotator agreement. Section 5 presents our first experiments using this dataset for neural machine learning of fine-grained opinions, before Section 6 discusses some future directions of research. Finally, Section 7 summarizes the main contributions of the paper.

2. Related Work

Fine-grained approaches to sentiment analysis include opinion annotations as in (Wiebe et al., 2005), aspect-based sentiment (Hu and Liu, 2004), and targeted sentiment (Vo and Zhang, 2015). Whereas document- and sentence-level sentiment analysis make the simplifying assumption that all polarity in the text is expressed towards a single entity, fine-grained approaches attempt to model the fact that polarity is directed towards entities (either implicitly or explicitly mentioned). In this section we provide a brief overview of related work, first in terms of datasets and then modeling.

2.1. Datasets

One of the earliest datasets for fine-grained opinion mining is the MPQA corpus (Wiebe et al., 2005), which contains annotations of private states in English-language texts taken from the news domain. The authors propose a detailed annotation scheme in which annotators identify subjective expressions, as well as their targets and holders.

Working with sentiment in English consumer reviews, Toprak et al. (2010) annotate targets, holders and polar expressions, in addition to modifiers like negation, intensifiers and diminishers. The intensity of the polarity is marked on a three-point scale (weak, average, strong). In addition to annotating explicit expressions of subjective opinions, Toprak et al. (2010) annotate *polar facts* that may *imply* an evaluative opinion. A similar annotation scheme is followed by Van de Kauter et al. (2015), working on financial news texts in Dutch and English, also taking account of implicit expressions of sentiment in polar facts.

The SemEval 2014 shared task (Pontiki et al., 2014) proposes a different annotation scheme. Given an English tweet, the annotators identify targets, the aspect category they belong to, and the polarity expressed towards the target. They do not annotate holders or polar expressions.

While most fine-grained sentiment datasets are in English, there are datasets available in several languages, such as German (Klinger and Cimiano, 2014), Czech (Steinberger et al., 2014), Arabic, Chinese, Dutch, French, Russian, Spanish, Turkish (Pontiki et al., 2016), Hungarian (Szabó et al., 2016), and Hindi (Akhtar et al., 2016). Addition-

---

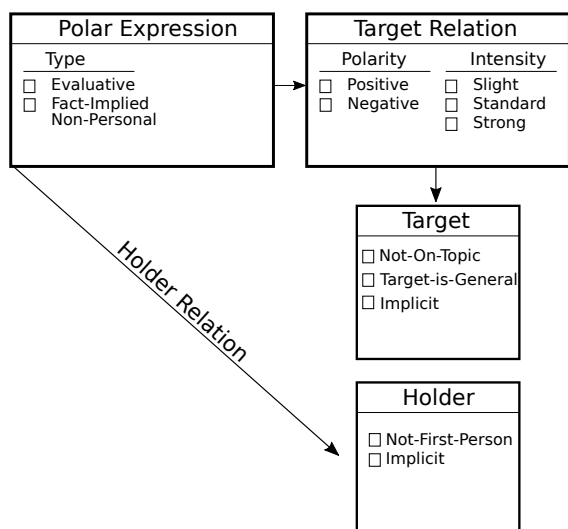[1] https://github.com/ltgoslo/norec_fine

Figure 1: Annotation schema for the NoReC_*fine* dataset.

Figure 2: Annotation of an EVAL sentence (transl. 'This disk runs very quietly').

ally, there has been an increased effort to create fine-grained resources for low-resource languages, such as Basque and Catalan (Barnes et al., 2018). No datasets for fine-grained SA have previously been created for Norwegian, however.

2.2. Modeling

Fine-grained sentiment is most often approached as a sequence labeling problem (Yang and Cardie, 2013; Vo and Zhang, 2015) or simplified to a classification problem when the target or aspect is given (Pontiki et al., 2014).

Although the specific architectures vary greatly, state-of-the-art methods for fine-grained sentiment analysis tend to be transfer-learning approaches (Chen and Qian, 2019), often using pre-trained language models (Peters et al., 2018; Devlin et al., 2019) to improve model performance (Hu et al., 2019). Additionally, several approaches use multi-task learning to incorporate useful information from similar tasks into fine-grained sentiment models (Marasović and Frank, 2018; He et al., 2019; Li et al., 2019).

Alternatively, researchers have proposed attention-based methods which are adapted for fine-grained sentiment (Tang et al., 2019; Bao et al., 2019). These methods make use of an attention mechanism (Bahdanau et al., 2014) which allows the model to learn a weighted representation of sentences with respect to sentiment targets.

Finally, there are approaches which create task-specific models for fine-grained sentiment. Liang et al. (2019) propose an aspect-specific gate to improve GRUs.

3. Annotations

In the following we present our fine-grained sentiment annotation effort in more detail. We provide an overview of the annotation guidelines and present statistics on inter-annotator agreement. The complete set of guidelines is distributed with the corpus.
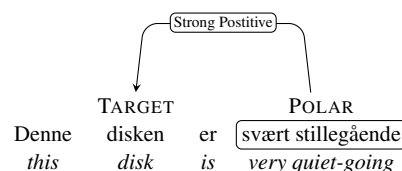
Sentence-level annotations We build on the sentence-level annotation of evaluative sentences in the NoReC_*eval*-corpus (Mæhlum et al., 2019), where two types of evaluative sentences were annotated: simple *evaluative* sentences (labeled EVAL), or the special case of *evaluative fact-implied non-personal* (FACT-NP) sentences. The 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, e.g., *elske* 'love', *like* 'like', *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', *verdt (penger, tid)* 'worth (money, time)'. In NoReC_*eval*, 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.

While previous work (Toprak et al., 2010) only annotate sentences that are found to be 'topic relevant', Mæhlum et al. (2019) choose to annotate all sentiment-bearing sentences, but explicitly include a Not-on-Topic marker. This will allow for assessing the ability of models to reliably identify sentences that are not relevant but still evaluative.

Expression-level annotations In our current fine-grained annotation effort we annotate both the EVAL and FACT-NP sentences from the NoReC_*eval* corpus. Figure 1 provides an overview of the annotation scheme and the entities, relations and attributes annotated. Example annotations are provided in Figure 2, for an EVAL sentence, and Figure 3 for a FACT-NP. As we can see, positive or negative polarity is expressed by a relation between a *polar expression* and the *target(s)* of this expression and is further specified for its strength on a three-point scale, resulting in six polarity values, ranging from *strong positive* to *strong negative*. The *holder* of the opinion is also annotated if it is explicitly mentioned. Some of the annotated entities are further annotated with attributes indicating, for instance, if the opinion is not *on topic* (in accordance with the topic of the review) or whether the target or holder is *implicit*.

3.1. Polar Expressions

A polar expression is the text span that contributes to the evaluative and polar nature of the sentence. For some sentences this may simply be expressed by a sentiment lexeme such as *elsker* 'loves', *forferdelig* 'awful' for EVAL type expressions. In the case of FACT-NP polar expressions,
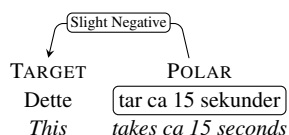
Figure 3: Annotation of a FACT-NP sentence (transl. 'This takes around 15 seconds').
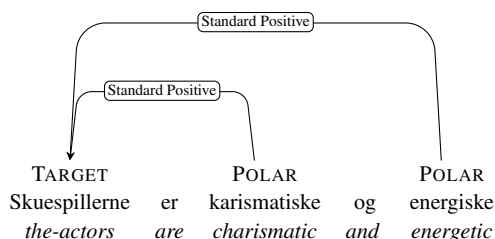

Figure 4: An example of coordinated polar expressions (transl. 'The actors are charismatic and energetic').

any objective description that is seen to reflect the holder's evaluation is chosen, as in Figure 3. Polar expressions may also include modifiers, including intensifiers such as *very* or modal elements such as *should*. Polar expressions are often adjectives, but verbs and nouns also frequently occur as polar expressions. In our annotation, the span of a polar expression should be large enough to capture all necessary information, without including irrelevant information. In order to judge what is relevant, annotators were asked to consider whether the strength and polarity of the expression would change if the span were reduced.

Polar expression span The annotation guidelines further describe a number of distinctions that should aid the annotator in determining the polar expression and its span. Certain punctuation marks, such as exclamation and question marks, can be used to modify the evaluative force of an expression, and are therefore included in the polar expression if this is the case.

Verbs are only included if they contribute to the semantics of the polar expression. For example, in the sentence in Figure 5 the verb *led* 'suffers' clearly contributes to the negative sentiment and is subsequently included in the span of the polar expression. High-frequent verbs like *å være* 'to be' and *å ha* 'to have' are generally not included in the polar expression, as shown in the example in Figure 2 above. Prepositions belonging to particle verbs and reflexive pronouns that occur with reflexive verbs are further included in the span. Verbs that signal the evaluation of the author but no polarity are not annotated, These verbs include *synes* 'think' and *mene* 'mean'.

Sentence-level adverbials such as *heldigvis* 'fortunately', *dessverre* 'unfortunately', often add evaluation and/or polarity to otherwise non-evaluative sentences. In our scheme, they are therefore annotated as part of the polar expression. Coordinated polar expressions are as a general rule treated as two separate expressions, as in the example in Figure 4 where there are two conjoined polar expressions with sep-

arate target relations to the target. In order to avoid multiple (unnecessary) discontinuous spans, conjunct expressions that share an element, are, however, included in the closest conjunct. An example of this is found in Figure 5, where the verbal construction *led av* 'suffered from' has both syntactic and semantic scope over both the conjuncts (*led av dårlig dialog* 'suffered from bad dialog' and *led av en del overspill* 'suffered from some over-play'). If the coordinated expression is a fixed expression involving a coordination, the whole expression should be marked as one coherent entity.

Expletive subjects are generally not included in the span of polar expressions. Furthermore, subjunctions should not be included unless excluding them alone leads to a discontinuous span.

Polar expression intensity The intensity of a polar expression is indicated linguistically in several different ways. Some expressions are inherently strongly positive or negative, such as *fabelaktig* 'fabulous', and *katastrofal* 'catastrophic'. In other cases, various modifying elements shift the intensity towards either point of the scale, such as adverbs, e.g., *uhyre* 'immensely' as in *uhyre tynt* 'immensely thin'. Some examples of adverbs found with slightly positive or negative expressions are *noe* 'somewhat', *kanskje* 'maybe' and *nok* 'probably'. The target of the expression can also influence the intensity, and the annotators were urged to consider the polar expressions in context.

3.2. Targets

We annotate the targets of polarity by explicitly marking target entities in the text and relating them to the corresponding polar expression via a target relation. In Figure 2 for instance we see that the polar expression *svært stillegående* 'very quiet-going' is directed at the target *disken* 'disk' and expresses a Strong Positive polarity. As a rule of thumb, the span of a target entity should be as short as possible whilst preserving all relevant information. This means that information that does not aid in identifying the target should not be included. When it comes to more formal properties of targets, they are typically nominal, but in theory they can also be expressed through adjectives or verbs. Target identification is not always straightforward. Our guidelines therefore describe several guiding principles, as well as some more detailed rules of annotation. For instance, reviewed objects might have easily identifiable physical targets, e.g., a tablet can have the targets *screen* and *memory*. However, targets may also have more abstract properties, such as *price* or *ease of use*. A target can also be a property or *aspect* of another target. Following the tablet example above, the target *screen* can have the sub-aspects *resolution*, *color quality*, etc. We can imagine an aspect tree, spanning both upwards and downwards from the object being reviewed.

Canonical targets Targets are only selected if they are considered *canonical*, meaning that they represent some commonly encountered feature of the object under review. For example, in the sentence *det var en god nerve i hovedmysteriet* 'there was a good nerve in the main mystery', the word *nerve* 'nerve' is not annotated as target, as it is not
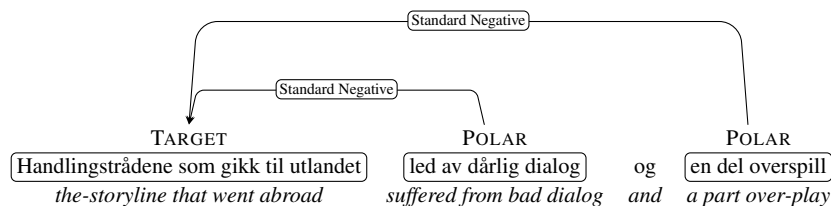
Figure 5: An annotated example of a coordinated polar expression (transl. 'The storyline that stretches abroad suffered from bad dialog and exaggerated acting').

considered to be an essential part of (in this case) all TV series in general. Rather, *god nerve* 'good nerve' is annotated as a polar expression targeted at the entity *hovedmysteriet* 'the main mystery'. Fixed expressions, such as *god idé* are a common source of non-canonical expressions. In contrast, in phrases such as *tydelig skuespill* 'clear acting', the target *skuespill* 'acting' is seen as an integral part of TV series, and therefore considered so-called canonical.

General targets When the polar expression concerns the object being reviewed, we add the attribute *target-is-general*. This applies both when the target is explicitly mentioned in the text and when it is implicit. The *target-is-general* attribute is not used when a polar expression has a target that is at a lower ontological level than the object being reviewed, as for instance, in the case of the tablet's screen, given our previous example.

Implicit targets A polar expression does not need to have an explicit target. Implicit targets are targets that do not appear in the same sentence as the polar expression it relates to. We identify three types of implicit targets in our scheme: (i) implicit not-on-topic targets, (ii) implicit general targets and, (iii) implicit canonical aspect targets. A polar expression that refers to something other than what is being reviewed, is marked as *not-on-topic*, even if the reference is implicit. For marking a polar expression that is about the object being reviewed in general, the *target-is-general* attribute is used.

Polar–target combinations There are several constructions where targets and polar expressions coincide. Like most Germanic languages, nominal compounding is highly productive in Norwegian and compounds are mostly written as one token. Adjective–noun compounds are fairly frequent and these may sometimes express both polar expression and target in one and the same token, e.g., *favorittfilm* 'favourite-movie'. Since our annotation does not operate over sub-word tokens, these types of examples are marked as polar expressions.

3.3. Holders

Holders of sentiment are not frequently expressed explicitly in our data, partly due to the genre of reviews, where the opinions expressed are generally assumed to be those of the author. When they do occur though, holders are commonly expressed as pronouns, but they can also be expressed as nouns such as *forfatteren* 'the author', proper names, etc. Figure 6 shows an annotated example where the holder of the opinion *Vi* 'We' is related to a polar expression. Note that this example also illustrates the treatment of discon-

Figure 6: An annotated example sentence with an explicit holder (transl. 'We liked the predecessor, Urbanite-XLi a lot').

tinuous polar expressions. Discontinuous entities are indicated using a dotted line, as in Figure 6 where the polar words *likte* 'liked' and *godt* 'well' form a discontinuous polar expression. At times, authors may bring up the opinions of others when reviewing, and in these cases the holder will be marked with the attribute Not-First-Person.
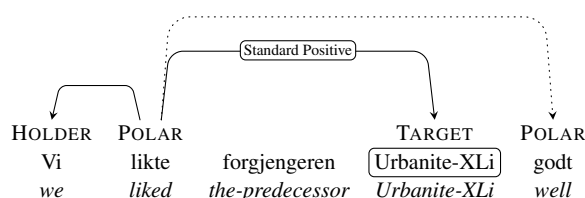
3.4. General

We will here discuss some general issues that are relevant for several of the annotated entities and relations.

Nesting In some cases, a polar expression and a target together form a polar expression directed at another target. If all targets in these cases are canonical, then the expressions are nested. Figure 7 shows an example sentence where the verb *ødelegger* 'destroys' expresses a negative polarity towards the target *spenningskurven* 'the tension curve' and the combination *ødelegger spenningskurven* 'destroys the tension curve' serves as a polar expression which predicates a negative polarity of the target *serien* 'the series'.

Comparatives Comparative sentences can pose certain challenges because they involve the same polar expression having relations to two different targets, usually (but not necessarily) with opposite polarities. Comparative sentences are indicated by the use of comparative adjectival forms, and commonly also by the use of the comparative subjunction *enn* 'than'. In comparative sentences like *X er bedre enn Y* 'X is better than Y', X and Y are entities, and *bedre* 'better' is the polar expression. In general we annotate *X er bedre* 'X is better' as a polar expression modifying Y, and *bedre enn Y* 'better than Y' as a polar expression modifying X. Here there should be a difference in polarity as well, indicating that X is better than Y. The annotated examples in Figure 8 shows the two layers of annotation invoked by a comparative sentence.
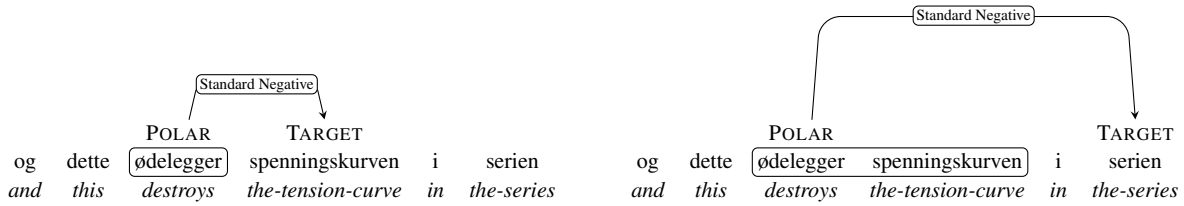
5028

Figure 7: Example of a sentence with nested annotation, each level of nesting is shown as a separate annotation layer (transl. 'and this destroys the plot development of the series').
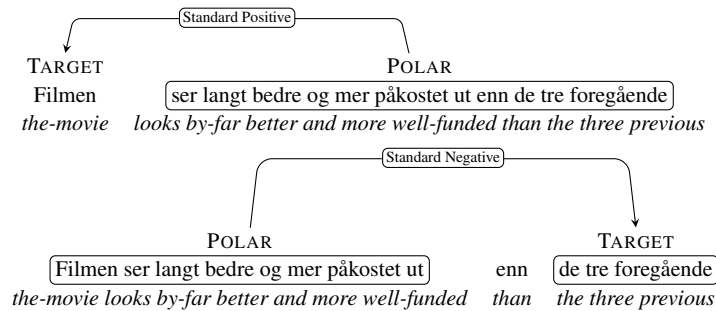

Figure 8: An example of a comparative sentence with nested annotations, here shown as two separate layers of annotation (transl. 'The movie looks far better and more expensive than the three previous ones').

Determiners Demonstratives and articles are generally not included in the span of any expressions, as exemplified by the demonstrative *Denne* 'this' in the example in Figure 2 above, unless they are needed to resolve ambiguity. Quantifiers such as *noen* 'some' , *mange* 'many' on the other hand are always included if they contribute to the polarity or intensity of the sentence.

3.5. Annotation Procedure

The annotation was performed by several hired student assistants with a background in linguistics and with Norwegian as their native language. All 298 documents in the dataset, comprising 7961 sentences, were annotated independently by two annotators in parallel. The doubly annotated documents were then adjudicated by a third annotator different from the two initial annotators. In the initial annotation phase, all annotators were given the possibility to discuss difficult choices in joint annotator meetings, but were encouraged to take independent decisions based on the guidelines if possible. Annotation was performed using the web-based annotation tool Brat (Stenetorp et al., 2012).

3.6. Inter-Annotator Agreement

In this section, we examine inter-annotator agreement. As extracting opinion holders, targets, and opinion expressions at token-level is a difficult task, even for humans (Wiebe et al., 2005), we use soft evaluation metrics, specifically *Binary Overlap* and *Proportional Overlap* (Katiyar and Cardie, 2016). Binary Overlap counts any overlapping predicted and gold span as correct. Proportional Overlap instead assigns precision as the ratio of overlap with the predicted span and recall as the ratio of overlap with the gold span, which reduces to token-level F_1. Proportional Overlap is therefore a stricter metric than Binary Overlap. Finally, we also compute the Cohen's kappa coefficient κ

for Proportional Overlap. We do not do the same for Binary Overlap, as the procedure leads to symmetrically imbalanced predictions (both annotator 1 and annotator 2 are highly likely to have overlapping annotations) and this is known to be problematic for κ (Flight and Julious, 2015; Zec et al., 2017). In order to calculate inter-annotator agreement for polarity and intensity labels, we report Binary Overlap and Cohen's κ on sentences which have a single polar expression. The reason for this restriction is to avoid having large spans from one annotator overlapping with several smaller spans from a second resulting in either artificially inflated or decreased agreement scores for polarity and intensity.

The inter annotator agreement scores obtained in the first rounds of (double) annotation are reported in Table 1. We find that even though annotators tend to agree on certain parts of the expressions, they agree less when it comes to exact spans. The proportional and binary holder scores are the highest (99% and 92% F_1). As mentioned earlier, holder expressions tend to be short, often pronominal, hence they are easier to agree on. Targets are more difficult, and may include longer expressions. Further, and as noted during annotation, there is strong agreement on the most central elements of the polar expression, even though there are certain disagreements when it comes to the exact span of a polar expression. When it comes to targets, however, there is considerably less agreement among the annotators (73% Binary F_1). More analysis is required in order to fully understand the reasons for this, but it is possible that the distinction between canonical and non-canonical targets has proven difficult to implement in practice. One additional source of disagreement might be the fact that the same target can be mentioned several times in the same sentence, and this might lead to conflicting choices between annotators.

	Label	F_1	κ
Prop.	Holder	99%	0.64
	Target	91%	0.57
	P. Exp.	77%	0.50
Binary	Holder	92%	n/a
	Target	73%	n/a
	P. Exp.	90%	n/a

Table 1: Inter-annotator agreement F_1-scores for holders, targets, and polar expressions. We report F_1-scores for *Proportional Overlap* (percentage of token-level overlap between annotations) and *Binary Overlap* (any overlap between annotations counts towards true positives), as well as Cohen's κ for the proportional scores.

	F_1	κ
Polarity	91%	0.82
Intensity	67%	0.28

Table 2: Inter-annotator agreement F_1 and κ scores for polarity and intensity given binary Polar expression overlap.

Table 2 presents inter-annotator agreement for polarity (positive or negative) and intensity (strong, standard and slight) separately. We find that whereas the general polarity shows fairly high agreement (91% F_1 and 0.82 κ), the annotation of intensity invokes considerably less agreement (67% F_1 and 0.28 κ) among the annotators.

4. Corpus Statistics

Table 3 presents some relevant statistics for the resulting NoReC$_{fine}$ dataset, providing the distribution of sentences, as well as holders, targets and polar expressions in the train, development and test portions of the dataset, as well as the total counts for the dataset as a whole. We also report the average length of the different annotated categories. As we can see, the total of 7961 sentences that are annotated comprise 7581 polar expressions, 5999 targets, and 735 holders. In the following we present and discuss some additional core statistics of the annotations.

Distribution of valence, intensity and polar expressions Figure 9 plots the distribution of polarity labels and their intensity scores. We see that the intensities are clearly dominated by *standard* strength, while there are also 1270 *strong* labels for positive. Regardless of intensity, we see that positive valence is more prominent than negative, and this reflects a similar skew for the document-level ratings in this data (Velldal et al., 2018).

The *slight* intensity is infrequent, with 190 positive and 330 negative polar expressions with this label. The relative difference here can be explained by the tendency to hedge negative statements more than positive ones (Kiritchenko and Mohammad, 2016). *Slight negative* is the minority class, with only 190 examples, followed by *strong negative* with 292 examples. Overall, the distribution of intensity

	# Examples				
	Train	Dev.	Test	Total	Avg. len.
Sents.	5915	1151	895	7961	16.8
Holders	584	76	75	735	1.1
Targets	4458	832	709	5999	2.0
Polar exp.	5659	1050	872	7581	4.6

Table 3: Number of sentences and annotated holders, targets and polar expressions across the data splits. The final column shows average token length. (Holders and targets can also be *implicit*, but these are not counted in this table.)
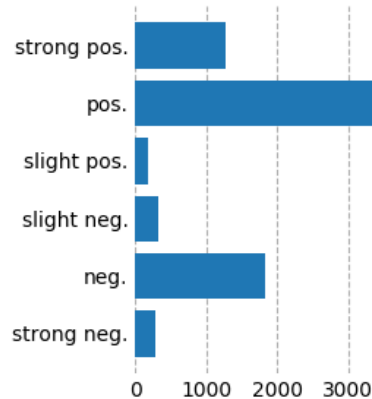

Figure 9: Distribution of labels and intensities.

scores in NoReC$_{fine}$ is very similar to what is reported for other fine-grained sentiment datasets for English and Dutch (Van de Kauter et al., 2015).

As we can see from Table 3, the average number of tokens spanned by a *polar expression* is 4.6. Interestingly, if we break this number down further, we find that the negative expressions are on average longer than the positives for all intensities: while the average length of negative expressions are 5.5, 5.4, and 5.7 tokens for *standard*, *strong*, and *slight* respectively, the corresponding counts for the positives are 4.2, 4.3, and 4.8. Overall, we see that the *slight* examples are the longest, often due to hedging strategies which include adverbial modifiers, e.g., 'a bit' or 'maybe'. Finally, note that only 380 of the annotated polar expressions are of the type *fact-implied non-personal*.

Distribution of holders and targets Returning to the token counts in Table 3, we see that while references to *holders* are just one word on average (often just a pronoun), *targets* are two on average. However, not all targets and holders have a surface realization. There are 6846 polar expressions with an *implicit holder*, 1426 with an *implicit target*, and 2186 with the tag *target-is-general*.

Finally, we note that there are 925 examples where the target is further marked as *not-on-topic* and 150 where the holder is *not-first-person*.

		P	R	F_1
Prop.	Holder	56.4	34.3	42.4
	Target	28.2	36.4	31.3
	P. Exp.	28.1	36.4	31.3
Binary	Holder	55.8	35.1	43.5
	Target	38.4	40.4	39.1
	P. Exp.	66.3	57.7	61.5

Table 4: Precision (P), recall (R), and Micro F_1 for holders, targets, and polar expressions. Prop. refers to proportional overlap and Binary refers to binary overlap.

5. Experiments

To provide an idea of the difficulty of the task, here we report some preliminary experimental results for the new dataset, intended as benchmarks for further experiments. Casting the problem as a sequence labeling task, we train a model to jointly predict holders, targets and polar expressions. Below, we first describe the evaluation metrics and the experimental setup, before finally discussing the results.

5.1. Experimental Setup

We train a Bidirectional LSTM with a CRF inference layer, which has shown to be competitive for several other sequence labeling tasks (Huang et al., 2015; Lample et al., 2016; Panchendrarajan and Amaresan, 2018). We use the IOB2 label encoding for sources, targets, and polar expressions, including the binary polarity of the latter, giving us nine tags in total.[2] This naturally leads to a lossy representation of the original data, as the relations, nested annotations, and polar intensity are ignored.

Our model uses a single BiLSTM layer (100 dim.) to extract features and then a CRF layer to make predictions. We train the model using Adam (Kingma and Ba, 2014) for 40 epochs with a patience of 5, and use dropout to regularize both the BiLSTM (0.5) and CRF (0.3) layers. The word embeddings are 100 dimensional fastText SkipGram (Bojanowski et al., 2016) vectors trained on the NoWaC corpus (Guevara, 2010) and made available from the NLPL vector repository (Fares et al., 2017).[3] The pre-trained embeddings are further fine-tuned during training. We report held-out test results for the model that achieves the best performance on the development set and use the standard train/development/test split provided with the dataset (shown in Table 3). All results are reported using the Proportional and Binary precision, recall and F_1 scores, computed as described in Section 3.6 above.

5.2. Results

Table 4 shows the results of the proportional and binary Overlap measures for precision, recall, and F_1. The baseline model achieves modest results when compared to

datasets that do not involve multiple domains (Yang and Cardie, 2013; Barnes et al., 2018), with 42.4, 31.3, and 31.3 Proportional F_1 on holders, targets, and polarity expressions, respectively (43.5, 39.1, 61.5 Binary F_1). However, this is still better than previous results on cross-domain datasets (Ding et al., 2017). The domain variation between documents leads to a lower overlap between holders, targets, and polar expressions seen in training and those at test time (56%, 28%, and 50%, respectively). We argue, however, that this is a more realistic situation regarding available data, and that it is important to move away from simplifications where training and test data are taken from the same distribution.

6. Future Work

In follow-up work we plan to further enrich the annotations with additional compositional information relevant to sentiment, most importantly negation but also other forms of valence shifters. Although our data already contains multiple domains, it is still all within the genre of reviews, and while we plan to test cross-domain effects within the existing data we would also like to add annotations for other different genres and text types, like editorials.

In terms of modeling, we also aim to investigate approaches that better integrate the various types of annotated information (targets, holders, polar expressions, and more) and the relations between them when making predictions, for example in the form of multi-task learning. Modeling techniques employing attention or aspect-specific gates that have provided state-of-the-art results for English provide an additional avenue for future experimentation.

7. Summary

This paper has introduced a new dataset for fine-grained sentiment analysis, the first such dataset available for Norwegian. The data, dubbed NoReC$_{fine}$, comprise a subset of documents in the Norwegian Review Corpus, a collection of professional reviews across multiple domains. The annotations mark polar expressions with positive/negative valence together with an intensity score, in addition to the holders and targets of the expressed opinion. Both subjective and objective expressions can be polar, and a special class of objective expressions called *fact-implied non-personal* expressions are given a separate label. The annotations also indicate whether holders are first-person (i.e., the author) and whether targets are on-topic. Beyond discussing the principles guiding the annotations and describing the resulting dataset, we have also presented a series of first classification results, providing benchmarks for further experiments. The dataset, including the annotation guidelines, are made publicly available.[4]

---

[2]The tagset is {O, B-Holder, I-Holder, B-Target, I-Target, B-Pos, I-Pos, B-Neg, I-Neg}.

[3]Corresponding to model ID 124 from the repository: http://vectors.nlpl.eu/repository/

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

Enstad, Anders Næss Evensen, Helen Ørn Gjerdrum, Petter Mæhlum, Lilja Charlotte Storset, Carina Thanh-Tam Truong, and Alexandra Wittemann.

Bibliographical References

Akhtar, M. S., Ekbal, A., and Bhattacharyya, P. (2016). Aspect based sentiment analysis in hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September.

Bao, L., Lambert, P., and Badia, T. (2019). Attention and lexicon regularized LSTM for aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 253–259, Florence, Italy, July.

Barnes, J., Badia, T., and Lambert, P. (2018). Multi-Booked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Chen, Z. and Qian, T. (2019). Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556, Florence, Italy, July.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota, June.

Ding, Y., Yu, J., and Jiang, J. (2017). Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *AAAI Conference on Artificial Intelligence*, pages 3436–3442.

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (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.

Flight, L. and Julious, S. A. (2015). The disagreeable behaviour of the kappa statistic. *Pharmaceutical Statistics*, 14(1):74–78.

Guevara, E. R. (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, June.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2019). An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy, July.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177, Seattle, USA.

Hu, M., Peng, Y., Huang, Z., Li, D., and Lv, Y. (2019). Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy, July.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Katiyar, A. and Cardie, C. (2016). Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany, August.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

Kiritchenko, S. and Mohammad, S. (2016). The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–52, San Diego, California, June. Association for Computational Linguistics.

Klinger, R. and Cimiano, P. (2014). The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Li, X., Bing, L., Li, P., and Lam, W. (2019). A unified model for opinion target extraction and target sentiment prediction. In *Proceedings the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019)*, pages 6714–6721, Honolulu, Hawaii, January. AAAI Press.

Liang, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., and Zhou, J. (2019). A novel aspect-guided deep transition model for aspect based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages –, Hong Kong, China, November.

Liu, B. (2015). *Sentiment analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press,

Cambridge, United Kingdom.

Mæhlum, P., Barnes, J., Øvrelid, L., and Velldal, E. (2019). Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Marasović, A. and Frank, A. (2018). SRL4ORL: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 583–594, New Orleans, Louisiana, June. Association for Computational Linguistics.

Panchendrarajan, R. and Amaresan, A. (2018). Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, June.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Steinberger, J., Brychcín, T., and Konkol, M. (2014). Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland, June. Association for Computational Linguistics.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Szabó, M. K., Vincze, V., Simkó, K. I., Varga, V., and Hangya, V. (2016). A hungarian sentiment corpus manually annotated at aspect level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Tang, J., Lu, Z., Su, J., Ge, Y., Song, L., Sun, L., and Luo, J. (2019). Progressive self-supervised attention learning for aspect-level sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 557–566, Florence, Italy, July.

Toprak, C., Jakob, N., and Gurevych, I. (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.

Van de Kauter, M., Desmet, B., and Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49:685–720.

Velldal, E., Øvrelid, L., Bergem, E. A., Stadsnes, C., Touileb, S., and Jørgensen, F. (2018). NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Vo, D.-T. and Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1347–1353, Buenos Aires, Argentina.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, May.

Yang, B. and Cardie, C. (2013). Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zec, S., Soriani, N., Comoretto, R., and Baldi, I. (2017). High agreement and high prevalence: The paradox of cohen's kappa. *The Open Nursing Journal*, 11:211–218.