A Diagnostic Dataset for Sentiment and Negation Modeling for Norwegian

Petter Mæhlum
Department of Informatics
University of Oslo
pettemae@ifi.uio.no

Erik Velldal
Department of Informatics
University of Oslo
erikve@ifi.uio.no

Lilja Øvrelid
Department of Informatics
University of Oslo
liljao@ifi.uio.no

Abstract

Negation constitutes a challenging phenomenon for many natural language processing tasks, such as sentiment analysis (SA). In this paper we investigate the relationship between negation and sentiment in the context of Norwegian professional reviews. The first part of this paper includes a corpus study which investigates how negation is tied to sentiment in this domain, based on existing annotations. In the second part, we introduce NoReC$_{NegSynth}$, a synthetically augmented test set for negation and sentiment, to allow for a more detailed analysis of the role of negation in current neural SA models. This diagnostic test set, containing both clausal and non-clausal negation, allows for analyzing and comparing the abilities of models to treat several different types of negation. We also present a case-study, applying several neural SA models to the diagnostic data.

1 Introduction

In sentiment analysis, negation is an instance of the more general category of valence shifters (Liu, 2015), i.e., expressions that modify the polarity of a sentiment expression. They can shift polarity entirely, or reduce or increase it, functioning as diminishers or intensifiers, respectively. Negation is a well-known challenge for the correct treatment of sentiment analysis (SA) related tasks, and several previous studies have discussed negation as a source of error for SA, such as Wiegand et al. (2010) and Barnes et al. (2019).

In this paper we present new data on the relationship between sentiment and negation in Norwegian, based on the Norwegian Review Corpus (NoReC) (Velldal et al., 2018), a collection of professional reviews spanning several different domains, where the same subset of documents have been annotated for both structured sentiment (NoReC$_{fine}$; Øvrelid et al., 2020a), and negation (NoReC$_{neg}$; Mæhlum et al., 2021). To be able to more precisely test the sensitivity of a model to the correlations between these two phenomena, we introduce a new diagnostic dataset, NoReC$_{\mathrm{NegSynth}}$,[1] which contains synthetically constructed minimal pairs that illustrate more fine-grained negation phenomena. The dataset is created by extending NoReC$_{fine}$ by negating existing sentences in the test set, and then annotating them for sentiment and negation. These new sentences have then been annotated for unnaturalness. We further illustrate the use of this synthetic dataset by reporting diagnostics for three models, showing the potential of our dataset.

This paper is organized as follows. In Section 2 we describe relevant previous research on sentiment and negation. In Section 3 we analyze the relationship between sentiment and negation as it can be found in the two original datasets, discussing relevant corpus statistics. In Section 4 we turn to describe the creation of the new diagnostic dataset. In Section 5 we briefly introduce the models used to showcase the diagnostic data, also discussing the results. We propose some directions for future work in Section 6, before concluding in Section 7.

2 Related Work

Liu (2015) provides a thorough description of the interactions of negation and sentiment, albeit under the category of valence shifters. He includes an in-depth discussion of the relationship between certain negators in English and their various functions in expressing different types of sentiment. Recently another negation diagnosis test set was published for natural language inference (Truong et al., 2022). The authors note the lack of attention

[1] Available at https://github.com/Tyriflis/NorNegSynth

given to subclausal negation. Although not specifically designed with this in mind, our dataset naturally contains both clausal and non-clausal negation. This allows us to investigate certain negation cues that are exclusively subclausal, but does not distinguish between clausal and non-clausal usages of the same cue.

Several studies have focused on how diagnostic datasets can be used for SA, including Barnes et al. (2019), who present a challenge dataset for SA where instances that several SA-models get wrong are annotated for a range of different phenomena, noting that negation is one of the phenomena that affects SA the most. Hazarika et al. (2022) who create diagnostic tests to analyze robustness in multimodal SA.

3 Corpus Study on Negation and Sentiment

While a synthetically created diagnostic dataset can give us detailed insight into the performance of different models, it does not necessarily tell us anything about the relationship between sentiment and negation in actual language use. In order get a better understanding of the possible interactions between these two phenomena, we perform a corpus study on the NoReC$_{fine}$ and NoReC$_{neg}$ datasets. This preliminary study helped inform both the creation of our synthetic dataset and later the interpretation of the results obtained when evaluating models on it.

3.1 Datasets

In the sentiment-annotated dataset NoReC$_{fine}$ (Øvrelid et al., 2020a), an *opinion* consists of a *holder*, a *target* and a *polar* expression, in addition to the associated *polarity* (positive/negative) and its *intensity* (slight/standard/strong).

In NoReC$_{neg}$, the same texts as in NoReC$_{fine}$ have been annotated for negation. A *negation* consists of a *cue*, which is the word that triggers or identifies negation, and its *scope* within the sentence. Affixal cues such as the negating prefix *u-* 'un-' are also annotated. Since the cue can be understood more specifically as the lexical item indicating negation, we will also use the term *negator* for expressions that indicate negations, corresponding to a cue in use. The annotation guidelines for NoReC$_{fine}$ specify that all elements affecting the polarity of an opinion are to be included in the scope of the polar expression. This leads to all cues of polarity-modifying negations to be included in the scope of polar expressions, as in eample 1, where *verken* is annotated as belonging to both *tynn* and *blikkboks-aktig* separately.

	Total		Positive		Negative	
	#	%	#	%	#	%
Polarity	1756	100.0	724	41.0	1032	69.3
Standard	1266	53.7	499	39.4	767	60.6
Slight	195	61.0	51	26.2	143	73.3
Strong	293	51.4	172	58.7	121	41.3

Table 1: Negation overlap counts for polarity and intensity. The percentages indicate the proportions, but do not add up to 100, as the same negation expression can overlap with several different polar expressions.

(1) *Lyden* *er* *verken* *tynn* *eller*
Sound.the is neither thin nor
blikkboks-aktig
tin-can-like
'The sound is neither thin nor tin-can-like.'

Basic statistics of the relation between the complete SA and the negation datasets are reported in Table 1, broken down according to polarity and intensity. We present counts for positive and negative opinions separately, as we have seen that the relevant distributions are not always equal. We observe that negation co-occurs more frequently with negative polarity than positive. When it comes to intensity, negation also co-occurs more often with strong positive expressions as well as slight negative expressions.

3.2 Annotating the Effect of Negation on Sentiment

The direct effect of negation on sentiment is not apparent from the presence of negation alone. In order to more precisely study these interactions, we therefore manually annotate this effect on the 171 sentences in the test set that contain both negation and sentiment.

(2) *[Det* *er]* ***ikke*** *[viktig]* .
It is not important
'It is not important'

Given a sentence that contains at least one negation and at least one polar expression, as in example 2, a mapping is annotated from each negation to one or more of the polar expressions, indicating whether they are affected by the negation or not and in what way. As noted before, negation cues should be within the span of a polar expression if they affect the polarity of the expression (Øvrelid et al., 2020a). In example 2, the cue *ikke* negates the polar expression *viktig*.

However, the reverse is not true. There are cases where a negation cue is within the scope of a polar expression, without actually affecting the polarity of that expression, as in Example (3) below. Here the affixally negated word *urolig* 'uneasy' is within the polar expression itself (indicated in bold), but it is not itself part of what creates negative polarity, which presumably is mainly the verb *utfordrer* 'challenges'.

(3) *Det er fint, men* ***utfordrer den indre***
it is nice, but challenges the inner
og urolige seksåringen i oss alle*.*
and un-calm six-year-old.the in us all.
'It is nice, but challenges the inner and uneasy six-year-old in us all'

The results of the annotations are reported in Table 2. 'No change' indicates no effect on neither the polarity or intensity of the polar expression. 'No change (Negated)' refers to cases where a negation cue scopes over or is a part of the polar expression but there is no change in polarity. 'No change (Not negated)' refers to cases where the scope of the negation is outside the span of the polar expression and there is no change in polarity. If we consider the six-point scale employed in the sentiment annotation of NoReC$_{fine}$, with strong negative expressions representing the lower end of the scale, and strong positive expressions the upper, a *reduction* indicates a polarity shift towards the lower part of the scale, and an *increase* indicates change towards the upper. We find that the majority of negations in this study turn polar expressions into more negative expressions (Reduction), but that a non-negligible number of negators also increase polarity towards the more positive end of the scale.

3.3 Negation Cues in the Data

Looking more closely at the cues, focusing on those that have more than 3 occurrences in polar expressions, we see that there are differences in how they typically affect polarity. Table 3 shows how often these cues affect polarity as a reduction or increase in positive/negative valence. Only cues that occur inside polar expressions are counted. Cues that have their scopes outside are left out.

Type of change	#	%
No change (Negated)	34	10.3
No change (Not negated)	155	46.8
Positive to Negative	80	40.0
Negative to Positive	52	26.0

Table 2: Counts and percentage-wise distribution of shifts in polarity and cases where there is no change, depending on whether the negation cue is negated or not.

	Change		
Cue	Reduction	Increase	None
ikke 'not'	40	25	8
uten 'without'	3	6	2
u- 'un-'	15	5	13
aldri 'never'	2	4	0
-løs '-less'	1	4	1
ingen 'none; no one'	3	3	1
mangle 'lack'	6	0	0

Table 3: Reduction, increase or no change in polarity or intensity for the cues with more than three occurrences in the 171 sentences with both polarity and negation.

Among potentially interesting patterns, we note that *ikke* 'not' seems to be especially associated with reduction. The same tendency may be noted for *mangle* 'lack', but the lower frequencies makes it difficult to generalize. Moreover, the affixal negation cue *u-* is common both as a reducing negator and with no change, which partially stems from the tendency for many sentence-level adverbials to contain this cue.

Finally we return to the whole dataset and look more closely at the distribution of cues in relation to polarity and intensity. In Table 4 we see the 8 most common cues in the dataset and their respective distributions. This allows us to see that for example *ikke* 'not' is used much more frequently in relation to the 'slight' inensity, or that *u-* 'un-' seems to be somewhat more associated with strong intensity, and that although not as strongly, *aldri* 'never' seem to be more associated with positive polarity.

4 Synthetic Data for Diagnostics of Negation and Sentiment Modeling

Based on the corpus study described above, we have gained more insight into the relation between negation and sentiment. We now turn to describe

Type	*ikke*	*u-*	*uten*	*-løs*	*ingen*	*aldri*	*mangle*	*unntak*
Positive	47.5%	18.5%	8.5%	4.8%	6.5%	5.9%	0.9%	1.4%
Negative	55.0%	17.9%	5.6%	4.8%	4.6%	1.9%	4.2%	0.3%
Standard	52.4%	16.3%	7.7%	4.5%	5.6%	3.7%	3.3%	1.0%
Slight	71.2%	14.1%	3.5%	2.0%	4.6%	1.0%	0.5%	0.5%
Strong	37.7%	28.5%	5.3%	8.2%	5.3%	4.6%	2.0%	0.0%

Table 4: The eight most common cues and their respective distributions in the whole dataset (train, test, dev). Note that a single cue can occur in several different polar expressions, so the percentages do not add up to 100%.

the creation of a synthetic diagnostic dataset for evaluation of sentiment analysis models that measures specifically the effect of negation.

The diagnostic dataset was created based on existing NoReC annotations, by manually augmenting the test set using cues found in NoReC$_{neg}$, or by removing already existing negation cues. The complete process was as follows: 1) We first extracted sentences containing at least one polar expression, before 2) manually inspecting to see which negation cues could be used to negate the polar expressions in the sentence. 3) If a negation was applicable, we ascertain whether the cue can be added without significant syntactic changes. 4) The new sentences were given a new sentence ID, and each newly negated polar expression was mapped to its corresponding polar expression in the original sentence. 5) Once the dataset was completed, each sentence was annotated for negation, polarity, intensity and naturalness. The process was identical for the removal of cues. Examples 4 to 7 below illustrates the process. Here the original sentence has no negation, and then the cues *ikke* 'not', *ingen* 'no;none;no one' and *på ingen måte* 'in no way' were used to negate it.

(4) *En fest for sansene .*
A party for the senses
' A party for the senses'

(5) *Ikke en fest for sansene*
Not a party for senses.the
'Not a party for the senses'

(6) *Ingen fest for sansene*
No party for senses.the
'No party for the senses'

(7) *På ingen måte en fest for sansene*
On no way a party for senses.the
'In no way a party for the senses'

One benefit of this method is that it allows us to align the synthetic negated sentences to the preexisting annotations to create minimal negation pairs. Furthermore, non-manual methods such as rule-based negation insertion is difficult, as negation is a complex phenomenon that relies not only on syntactic constraints, but also on semantics. Not all polar expressions can be negated, and not all negations can be used in all cases. The large number of existence negators such as *strippe* 'stripped;bare',*fravær* 'absence', *savne* 'miss' and *blotte* 'void (of)' understandably have limited use, as they require an existing existential expression to negate. Another example is the negator *la være* 'refrain from', which requires a verb with an agentive subject, vastly restricting its distribution. In the synthetic sentence in Example 8, the original verb *gjort* 'done' has been substituted by *latt være å gjøre* 'refrained from doing', as *gjøre* 'do' allows for agentive subjects. Examples 4 to 7 are good examples of some of these restrictions. We see that the original sentence in example 4 cannot be negated with an existential negator, as there is no expression of existance to negate. The elided copula also does not take an agentive subject, leaving out possible negation with *la være* 'refrain from'.

(8) *[...] har alle latt være å gjøre en*
[...] have all leave be to done an
imponerende jobb med å gi hjerte og
impressive job with to give heart and
sjel til denne filmen .
soul to this movie .
' [...] have all done an impressive job with giving heart and soul to this movie.'

Another problem arises when a polar expression restricts the context in a way that hinders a negation from sounding natural. Expressions such as *ulempen er at ...*, 'The disadvantage is that ...' and *Heldigvis ...* 'Fortunately' require a polar expression to have negative and positive polarity, respectively, in order to sound natural.

4.1 Dataset Cues

NoReC$_{neg}$ contains a large number of cues that could potentially be interesting to investigate, but as discussed, not all cues allow insertion into any sentence. The cues found in the original dataset are reported alongside the number of increases in the synthetic dataset in Table 6. Frequency lists provide a good indicator of versatility; almost all negation cues can be rewritten with *ikke* 'not', but become increasingly specialized. The cues used in the synthetic dataset represent the most frequent cues found in NoReC$_{neg}$. We see that the cues *verken* and *ingen måte* constitute the largest differences compared to the original dataset. In order to avoid a high proportion of simple-to-use cues, the annotators were allowed to not annotate sentences using the most common cues, in favor of focusing on producing negations with less common ones.

4.2 Challenges

An attempt at trying to negate every sentence with every cue poses several challenges.

Cue limitations First of all, all cues have their own limitations. While *ikke*, being the most common negator, has a wide range of possible uses, the same cannot be said for e.g. *mangle*, 'lack', which requires an existential expression, and the beforementioned *la være*, or even *u-*, 'un-', which despite being the second-most frequent cue, is restricted to adjectives [2].

Embedding Expressions Another challenge is that the nature of the original sentences can make it difficult to construct natural-sounding examples. In some cases, the polar expressions we wish to negate are embedded in an overarching expression. Expressions such as *Ulempen er at* 'the catch is that', *Heldigvis* 'Fortunately' or *Det positive er at* 'The positive is that' already dictate the polarity of the following embedded phrase, and while it would have been interesting to investigate which effect this could have had, these negations lead to unnatural-sounding sentences, and were avoided.

4.3 Unnaturalness

One key point of our dataset is that it is similar to language that is likely to be found when working with review data. We wanted the same type of language. However, as discussed above, not all sentences can be negated in all ways, and negating might lead to unnatural-sounding sentences. In order to keep the dataset as natural-sounding as possible, an annotator separate from the creator of the synthetic dataset annotated all 306 sentences using an unnaturalness scale from 1–3, where 1 indicates that the annotator feels that they could produce the sentence in question themselves (low unnaturalness), 2 indicates that they do not find it strange, but they would not produce it themselves. Finally a score of 3 indicates that the sentence seems completely unnatural (high unnaturalness) . We found that 289 (95%) of the sentences are natural-sounding, while 13 (4%) were less natural. The 4 (1%) sentences receiving a naturalness score of 3 were all discarded.

	Positive		Negative	
Intensity	#	%	#	%
Slight	10	2.07	20	4.13
Standard	42	8.68	203	41.94
Strong	12	2.48	107	22.11
Total	64	16.24	330	83.76

Table 5: Polarity and intensity for polar expressions in the synthetic part of the dataset, given a change in polarity. The remaining 90 polar expressions had the same polarity as their original sentences.

4.4 Polarity and Intensity

The dataset was also annotated for polarity and intensity. The annotator was familiar with polarity annotation, but was asked to base the new annotation on the assumption that the existing annotation were correct. This was to more correctly annotate the effect of negation, rather than introduce new interpretations of the sentences. Out of the 302 natural-sounding sentences there was a total of 394 polar expression that had a change in polarity, while 90 polar expression kept their original polarity. The polarity and intensity of the changed expressions are reported in Table 5

4.5 Corpus Statistics

With the addition of the synthetic sentences, the combined test set contains 472 sentences with both negation and sentiment.

5 Benchmarking

In order to illustrate the use of our diagnostic dataset in the analysis of sentiment models we

[2] Note, this is not the denominal prefix *u-*, meaning 'bad'

Type	Count	Type	Count
ikke	158 + 156	nei	1 + 0
u	57 + 29	miste	1 + 0
uten	25 + 13	null	1 + 0
ingen	16 + 27	blotte	1 + 1
aldri	11 + 6	istedenfor	1 + 0
løs	10 + 1	strippe	1 + 0
mangle	6 + 11	ei	1 + 0
fravær	2 + 0	ingenting	1 + 1
ingen måte	2 + 21	mangel	1 + 0
unntak	2 + 0	la være	0 + 11
verken	2 + 25	unngå	0 + 4
fri	2 + 0	ikke-	0 + 3
savne	1 + 0		

Table 6: Cue counts in the original dataset and the added number in the artificial dataset. 23 sentences had negation removed.

apply three different models to it. The two first models are the baseline models for the 2022 SemEval shared task on Structured Sentiment Analysis (Barnes et al., 2022), while the third is a graph model presented in (Samuel et al., 2022). For all models we focus on polar expressions and the effect the different types of negation has on the interpretation of these if they are in the scope of a cue. Although some of the models have the capacity to treat both targets and holders in addition to polar expressions, we have chosen to ignore these expressions for the purpose of this dataset.

5.1 Sequence Labeling Model

The original SemEval 2022 baseline sequence labeling model (Barnes et al., 2022) employs the BIO tag scheme to mark polar expressions, targets and holders. The model originally first trains a separate BiLSTM model for each of the three parts, but in our case it was only trained on the polar expressions. After this, a relation prediction model is trained with another BilSTM with max pooling. The input words are represented by static embeddings; in our case those from Norwegian-Bokmaal CoNLL17 corpus.[3] Since this paper focuses on the effect of negation on polar expressions only, we only ran the model on the polar expressions in the dataset. It does not predict intensity.

5.2 Sentiment Graph Parser

The second baseline model (Barnes et al., 2021) employs a more advanced architecture. The underlying theory for this model is that a sentiment expression can be expressed as a graph, where the polar expression head is the root of the graph, hence reformulating the task to general graph parsing. The system is in essence a reimplementation of Dozat and Manning (2018). The model was run with static embeddings, which are the same as for the model above. The model architecture allows for the specification of whether the graphs should be *head-final*, indicating that the final token of an expression is the head, or *head-first*, where the first token indicates the head. In their original paper (Barnes et al., 2021), the best results were obtained with a head-final architecture, and we will only be using this in our evaluation scheme.

Model	Binary F_1	Token F_1	Pol. F_1
Seq. model	0.85	0.52	0.71
Graph parser	0.84	0.55	0.72
Dir. parser	0.79	0.56	0.87

Table 7: Binary, and tokens based F_1 scores for the three models. Polarity is only evaluated at the token level.

5.3 Direct Parsing Model

This model represents a near-state-of-the-art (SOTA) model which has showed good results for SA for Norwegian (Samuel et al., 2022). The model is also graph-based, but using a non-sequential semantic representation. It is an optimized version based on the graph parser presented in Samuel and Straka (2020), using contextualized embeddings.

5.4 Results

The output from running the three models on the dataset were evaluated separately for each model. From each model, the set of polar expressions with associated polarities were evaluted against the gold diagnostic dataset. We first evaluated each model output with a F_1-score in two granularities, token-based and binary, as in NoReC$_{fine}$ Øvrelid et al. (2020b), in order to get an overview of the models' capabilities. The results are shown in Table 7. Although the token F_1 is a better overall metric for the models, we use binary overlap to indicate matching polar expressions, as this disallows expressions lengths to influence the results.

From the results in Table 7, we see that although not indicated by the binary score, the direct parser scores better at token based F_1, and much more so when it comes to polarity. It is also worth mention-

[3] From the NLP Repository: `http://vectors.nlpl.eu/repository/`

ing that the good scores of this model also comes from its ability regarding holders and targets of polar expression, which we will not investigate. We stress that the aim of NoReC$_{\text{NegSynth}}$is not to evalute models overall; only to give indications of their treatment of negation in relation with sentiment.

Having a brief understanding of the models capabilities, and knowing that they all manage to capture polar expressions to some extent, we move to the negation. We here assess each polar expression in the diagnostic set, seeing if it has been predicted, as defined by a binary overlap with a predicted polar expression. We then examine negations inside this expression, as well as polarity and intensity, and the cue lemma.

From table 8, we see the results of the evaluation of where the models agree with the gold diagnostic set, and how these predictions are distributed among the various cues in the expressions they predict. As the models need to correctly give the correct interpretation to the cues when assigning polarity, this information allows for a more fine-grained cue-oriented diagnosis. While most of the cues have far too few occurrences, we notice especially the 10 most common cues: *ikke* 'not', *u-* 'un-', *ingen* 'no;no-one', *mangle* 'lack', *verken* 'neither', *uten* 'without', *aldri* 'never', *ingen måte* 'no way', *la være* 'refrain (from)' and *-løs* '-free'. The interesting parts are the different agreement proportions observed for the models. For example, we see that the direct parser agrees with the gold set in almost three times as many cases when the negation is *ikke*. We see similar tendencies for the same model and the other cues, with two interesting exceptions: *verken* is very slightly associated with disagreement, indicating that this model might struggle more with this cue. We further find that for the affixal cue *u-*, the direct parsing model is excelent at correctly identifying the correct sentiment, potentially indicating that this is a stronger model for subtoken negation. For the sequence tagger we see that the tendencies are lower than for the two other models. Although *u*, *mangle* and *uten* tend towards agreement, the others do not. We note that *verken* has a very low ratio here. The graph model places itself in between the two others. It tends towards agreement for all the top cues except *verken*, but not to the same extent as the direct parser.

5.5 Minimal Negation Pairs

As the negations were added to existing sentences with polarity, the resulting dataset contains several sets of *minimal pairs*, where the only difference is negation. Looking back at example **??** to example 7 , we see that the first example is the originally unnegated sentence. The three following sentences are negations using *ikke* 'not', *ingen* 'no' and *på ingen måte* 'in no way', respectively. The two first negations were annotated as Negative Standard, while the last was annotated with Negative Strong, which is typical of expressions with *på ingen måte*.

We theorize that looking at these sentences gives us more information about a models' negation analyzing capabitilies, in the sense that if it correctly predicts polar expressions in the original sentence, and also in the negated sentence, given that the polar expression in question is in fact negated by the newly added negation, then this must indicate that the model can interpret the cue correctly. There are 203 such pairs in the dataset. Most pairs (134) consist of only the original sentence and a single negatated sentence, but there are also sets with two negated (56), 3 negated (22) and a single case of four negated sentences for the same original sentence.

Among these sentences, not all possible polar expressions have been identified by all models. In order to be inspected, the original non-negated polar expression must have been predicted with binary overlap, along with at least one negated polar expression. We use this overlap to see how well the models are able to correctly identify negation and the related polarity change for these minimal pairs.

In Table 9 we see to which degree the three models are able to correctly predict these minimal pairs. We observe that in fact the Direct parser model and the Sequence model are both outperformed by the Graph model when it comes to correctly predicting the shift in polarity when adding negation to an originally Positive sentence. However, when adding negation to originally Negative sentences, we see that the Direct parser outperforms the other models. This comparison would not have been clear without these negation pairs.

6 Future Work

Despite being able to shed some light on the effects of various models, there is still the problem of scarce cues. If the goal is to maintain a high level of naturalness in the sentences, one possible

	Direct parser			Sequence			Graph		
Cue	Agree	Dis.	Ratio	Agree	Dis.	Ratio	Agree	Dis.	Ratio
ikke	66	23	2.87	46	59	0.78	60	45	1.33
u	31	4	7.75	26	14	1.86	21	15	1.40
ingen	12	2	6.00	6	11	0.55	9	7	1.29
mangle	10	4	2.50	13	2	6.50	10	5	2.00
verken	6	7	0.86	5	13	0.38	5	13	0.38
uten	9	0	-	8	5	1.60	10	4	2.50
aldri	3	1	3.00	0	5	-	3	2	1.50
ingen måte	6	2	3.00	2	7	0.29	9	4	2.25
la være	4	1	4.00	3	1	3.00	2	4	0.50
løs	3	2	1.50	4	2	2.00	4	3	1.33
blotte	2	2	1.00	0	0	-	2	2	1.00
savne	1	0	-	1	0	-	0	1	-
unngå	0	0	-	0	0	-	0	1	-
ingenting	0	0	-	0	1	-	1	0	-
ei	0	1	-	0	1	-	1	0	-
fri	2	0	-	1	0	-	1	1	1.00
fravær	0	0	-	1	1	1.00	0	1	-
ikke-	2	1	2.00	0	2	-	0	2	-
strippe	0	1	-	1	0	-	0	1	-

Table 8: Agreement and disagreement on polarity for the 19 negation cues in the corpus, for each of the three models.

	Direct		Graph		Seq	
Type	#	%	#	%	#	%
P to N (w)	100	60%	67	43%	112	70%
P to N (c)	40	24%	70	45%	20	13%
N to P (w)	10	6%	12	8%	24	15%
N to P (c)	18	11%	7	4%	3	2%
Total	168		156		159	

Table 9: How well each model is able to correctly (c) or incorrectly (w) predict the polarity of a negated sentence given that it correctly predicts its non-negated counterpart, from Positive (P) to Negative (N) and vice versa.

solution might be to actively seek out specific cues in the original un-annotated dataset and annotate them for polarity, rather than the opposite, as we have done here. As our research has shown that there does seem to be differences between different cues and how models treat them, we urge the exploration of individual cues and expressions to an even larger degree, especially those with low frequencies. Work also remains to explore how these tendencies generalize across different domains.

7 Conclusion

We have performed basic statistic checks of the relationship between negation and sentiment in Norwegian review texts and seen cases where certain negators co-occur more frequently with certain types of polarity and intensity. This motivated the creation of a synthetic dataset, which was used to evaluate three models. We see that this dataset reveals differences in how different machine learning models treat different cues, and how they differ in their ability to correctly identify polarity in minimal negation pairs, when the non-negated sentence is correctly identified. Furthermore, we see that although it is still difficult to include low-frequency cues, due to their limited syntactic and semantic flexibility, the increased number of common cues allow us to observe differences with greater confidence. We also note that this allows us to investigate cues that were not present in the original test set. We observe that having minimal negation pairs allows us to gain insight into the capabilities of the model which would not have been possible without these data. As the number of cues in the diagnostic dataset is comparable to the full test set, we believe that this type of diagnostic set can also save future annotation efforts where applicable.

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 would like to thank Anders Næss Evensen for annotating.

References

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. *CoRR*, abs/2105.14504.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States. Association for Computational Linguistics.

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

Petter Mæhlum, Jeremy Barnes, Robin Kurtz, Lilja Øvrelid, and Erik Velldal. 2021. Negation in Norwegian: an annotated dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 299–308, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020a. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference*, Marseille, France.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020b. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–478, Dublin, Ireland. Association for Computational Linguistics.

David Samuel and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The nan-nli test suite for sub-clausal negation.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, 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.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala, Sweden. University of Antwerp.