# Lexicon information in neural sentiment analysis: a multi-task learning approach

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

## Abstract

This paper explores the use of multi-task learning (MTL) for incorporating external knowledge in neural models. Specifically, we show how MTL can enable a BiLSTM sentiment classifier to incorporate information from sentiment lexicons. Our MTL set-up is shown to improve model performance (compared to a single-task set-up) on both English and Norwegian sentence-level sentiment datasets. The paper also introduces a new sentiment lexicon for Norwegian.

## 1 Introduction

Current state-of-the-art neural approaches to sentiment analysis tend not to incorporate available sources of external knowledge, such as polarity lexicons (Hu and Liu, 2004; Taboada et al., 2006; Mohammad and Turney, 2013), explicit negation annotated data (Morante and Daelemans, 2012; Konstantinova et al., 2012), or labels representing inter-annotator agreement (Plank et al., 2014). One reason for this is that neural models can already achieve good performance, even if they only use word embeddings given as input, as they are able to learn task-specific information (which words convey sentiment, how to resolve negation, how to resolve intensification) in a data-driven manner (Socher et al., 2013; Irsoy and Cardie, 2014). Another often overlooked reason is that it is not always entirely straightforward how we can efficiently incorporate this available external knowledge in the model.

Despite achieving strong results, neural models are known to be difficult to interpret, as well as highly dependent on the training data. Resources like sentiment lexicons, on the other hand, have the benefit of being completely transparent, as well as being easy to adapt or update. Additionally, lexicons are often less sensitive to domain and frequency effects and can provide high coverage and precision even for rare words. We hypothesize that these two views of sentiment are complimentary and that even competitive neural models can benefit from incorporating lexicon information.

In the current work, we demonstrate that multi-task learning (Caruana, 1993; Collobert et al., 2011) is a viable framework to incorporate lexicon information in a sentence-level sentiment classifier. Our proposed multi-task model shares the lower layers in a multi-layer neural network, while allowing the higher layers to adapt to either the main or auxiliary task. Specifically, the shared lower layers are a feed-forward network which uses a sentiment lexicon auxiliary task to learn to predict token-level sentiment. The higher layers use these learned representations as input for a BiLSTM sentiment model, which is trained on the main task of sentence-level classification. The intuition is that the representations learned from the auxiliary task give the model an advantage on the main task.

Compared to previous methods, our model has two advantages: 1) it requires *only a single sentiment lexicon*, and 2) the lexicon prediction model is *able to generalize to words that are not found in the lexicon*, increasing the overall performance. Experimental results are reported for both English and Norwegian, with the code[1] made available. While we rely on an existing sentiment lexicon for English, we introduce and make available a new lexicon for Norwegian.[2]

In the following, we first consider relevant related work (§ 2), and describe the sentiment lexicons (§ 3) and datasets (§ 4) that we use for our experiments. In § 5 we detail our proposed multi-task model, while § 6 presents the experimental results and error analysis. Finally, we summarize and point to future directions in § 7.

[1] https://github.com/ltgoslo/mtl_lex
[2] https://github.com/ltgoslo/norsentlex



*Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 175–186
Turku, Finland, 30 September – 2 October, 

## 2 Related work

In this section we briefly review previous relevant work related to (*i*) sentiment lexicons, (*ii*) lexicon-based approaches to sentiment analysis (SA), (*iii*) use of lexicon information in neural models, and finally (*iv*) multi-task learning in NLP.

**Sentiment lexicons** Sentiment lexicons provide a valuable source of information about the prior affective orientation of words, oftentimes driven by theoretical approaches to emotion (Stone et al., 1962; Bradley et al., 1999). There are several freely available sentiment lexicons for English. One widely used lexicon is that of Hu and Liu (2004),[3] which was created using a bootstrapping approach from WordNet and a corpus of product reviews. This is the lexicon that forms the basis of the experiments in the current paper and we return to it in § 3.1. Other available lexicons include the MPQA subjectivity lexicon (Wilson et al., 2005) which contains words and expressions manually annotated as positive, negative, both, or neutral. SentiWordnet (Esuli and Sebastiani, 2006) contains each synset of the English WordNet annotated with scores representing the sentiment orientation as being positive, negative, or objective. The So-Cal (Taboada et al., 2011) English sentiment lexicon contains separate lexicons of verbs, nouns, adjectives, and adverbs. The words were manually labeled on a scale from extremely positive ($+5$) to extremely negative ($-5$), and all words labeled as neutral ($0$) were excluded from the lexicons.

While no high-quality sentiment lexicons for Norwegian are currently publicly available, there have been some previous attempts at generating lexicons for Norwegian. Hammer et al. (2014) used a set of 51 positive and 57 negative manually selected seed words to crawl three Norwegian thesauri in three iterations, to extract synonyms and antonyms at each iteration. These were thereafter used to build an undirected graph with words as nodes, and synonymy and antonymy relations as edges. A label propagation algorithm was applied to create a lexicon by identifying the strength and polarity of the non-seed words and calculating the weighted average of the connected nodes. They used the Norwegian full-form lexicon SCARRIE[4] to retrieve all forms of each word in the lexicon. As a benchmark, they have also created two other lexicons: a machine translated version of the AFINN lexicon (Nielsen, 2011), and a manually corrected version of this translation. The generated lexicons were evaluated against reviews containing ratings (dice values) by summing the scores of each sentiment word present in a review, averaging these scores over the total number of words in the review, and assigning a final score based on threshold intervals. The authors also took into account the use of the sentiment shifter *ikke* (not) if it appeared two words before a word from the lexicons. Their results show that the translated lexicons outperformed, by mean absolute error with standard deviation, all of their automatically generated lexicons. Unfortunately, none of the lexicons are made available.

Bai et al. (2014) used a corpus of newspaper articles and discussion forums with a modified version of Pointwise Mutual Information (PMI) to compute the semantic orientation of candidate words against a list of seed words. They manually selected 7 positive and 7 negative words as seed words, and instead of using the entire corpus as candidate words they used a selection of the top 10,000 most frequent words in the corpus and a list of adjectives generated from their corpus using SCARRIE. Their results showed that the translated lexicons outperformed all of their generated lexicons, but unfortunately only the latter were made publicly available.

**Lexicon-based approaches to SA** Early approaches to sentiment analysis classified documents based on the sum of *semantic orientation* scores of adjectives in a document. Often, researchers used existing lexicons (Stone et al., 1962), or extended these resources in a semi-supervised fashion, using WordNet (Hu and Liu, 2004; Kim and Hovy, 2004; Esuli and Sebastiani, 2006). Alternatively, an adjective's semantic orientation could be determined as the strength of association with positive words (*excellent*) or negative words (*poor*) as measured by Pointwise Mutual Information (Turney and Littman, 2003).

Researchers quickly discovered, however, that various linguistic phenomena, *e.g.* negation, intensifying adverbs, downtoners, *etc*, must be taken into account to correctly assign a sentiment score. Taboada et al. (2011) proposed an approach to determine the semantic orientation of documents which incorporates sentiment lexicons for adjec-

[3] Available at `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

[4] `https://www.nb.no/sprakbanken/show?serial=sbr-9`

tives, nouns, verbs, and adverbs. Additionally, they included compositional rules for intensification, negation, and irrealis blocking. They showed that smaller, manually built lexicons outperform semi-supervised lexicon approaches and that their model is more robust to domain shifts than machine learning models.

**Lexicons in neural approaches** The general tendency in NLP when using neural approaches is to perform end-to-end learning without using external knowledge sources, relying instead solely on what can be inferred from (often pre-trained) word embeddings and the training corpus itself. This is also the case for neural sentiment modeling. However, there have been some attempts to include external knowledge like lexicon features into such models (Teng et al., 2016; Zou et al., 2018; Lei et al., 2018a; Bao et al., 2019a).

One notable example is the work of Shin et al. (2017) where several approaches are tested for how to incorporate lexicon information into a CNN for sentiment classification on the SemEval 2016 Task 4 dataset and the Stanford Sentiment Treebank (SST). Shin et al. (2017) create feature vectors that encode the positive or negative polarity values of words across a broad selection of different sentiment lexicons available for English. These word-level sentiment-score vectors are then combined with standard word embeddings in different ways in the CNN: through simple concatenation, using multiple channels, or performing separate convolutions. While all three approaches yield improvements for the SemEval data, performance deteriorates or remain unchanged for SST. The model used by Shin et al. (2017) requires information from six different lexicons, which is overly restrictive for most languages besides English, where one will typically not have the luxury of several publicly available sentiment lexicons.

Lei et al. (2018b) propose a different approach based on what they dub a 'Multi-sentiment-resource Enhanced Attention Network', where lexicon information is used for guiding an attention mechanism when learning sentiment-specific sentence representations. The approach shows promising results on both SST and the Movie Review data of Pang and Lee (2005), although the model also incorporates other types of lexicons, like negation cues and intensifiers.

In a similar spirit, Margatina et al. (2019) include features from a range of sentiment-related lexicons for guiding the self-attention mechanism in an LSTM. Bao et al. (2019b) generate features from several different lexicons that are added to an attention-based LSTM for aspect-based sentiment analysis.

In the current paper we will instead explore whether lexicon information can be incorporated into neural models using the framework of *multi-task learning*. This has two main advantages: 1) we require only a single sentiment lexicon, unlike much previous work, and 2) our model is able to generalize to sentiment words not seen in the lexicon as it only uses word embeddings as features. Below we review some relevant background on multi-task learning for NLP.

**Multi-task learning** Multi-task learning (MTL) (Caruana, 1993; Collobert et al., 2011), whereby a single machine learning model is simultaneously trained to perform two or more tasks, can allow a model to incorporate a useful inductive bias by restricting the search space of possible representations to those that are predictive for both tasks. MTL assumes that features that are useful for a certain task should also be predictive for similar tasks, and in this sense effectively acts as a regularizer, as it prevents the weights from adapting too much to a single task.

The simplest approach to MTL, *hard parameter sharing* (Caruana, 1993), assumes that all layers are shared between tasks except for the final predictive layer. This approach tends to improve performance when the auxiliary task is carefully chosen (Plank, 2016; Peng and Dredze, 2017; Martínez Alonso and Plank, 2017; Fares et al., 2018; Augenstein et al., 2018). What characteristics determine a useful auxiliary task, however, is still not completely clear (Bingel and Søgaard, 2017; Augenstein and Søgaard, 2017; Martínez Alonso and Plank, 2017; Bjerva, 2017).

Søgaard and Goldberg (2016) propose an improvement over hard parameter sharing that uses the lower layers of a multi-layer recurrent neural network to make predictions for low-level auxiliary tasks, while allowing higher layers to focus on the main task. In this work, we adopt a similar approach to incorporate sentiment lexicon information as an auxiliary task to improve sentence-level sentiment and evaluative-language classification.

## 3 Sentiment lexicons

We here describe the sentiment lexicons used in the experiments reported in § 5.

### 3.1 English sentiment lexicon

For English we use the sentiment lexicon compiled by Hu and Liu (2004), containing 4,783 negative words and 2,006 positive words. The sentiment lexicon was a bi-product of their task for predicting which reviews were positive and negative from a corpus of customer reviews of a range of products. Hu and Liu (2004) first PoS-tagged the review corpus to identify all the adjectives it contained, and then manually defined a list of 30 seed adjectives and their labels (positive or negative). The synsets and antonyms of the adjectives in the seed list were searched for in WordNet, and the positive and negative labels were automatically assigned based on the synonymy or antonymy relation of each word to the corresponding adjective from the seed list, iteratively growing the set of words in the lexicon. This has also enabled the inclusion of words that were not adjectives, which made the lexicon a mix of word classes and inflections.

### 3.2 Norwegian sentiment lexicon

We automatically translated (from English to Norwegian) the positive and negative words in the sentiment lexicon compiled by Hu and Liu (2004) described above. Thereafter, all the translations were manually inspected and corrected when necessary. If an English word had several senses that could be translated into different Norwegian words, these were manually added to the translations during the manual inspection. For example the English word *outstandingly* has been translated to the five following Norwegian words *bemerkelsesverdig*, *fortreffelig*, *fremstående*, *utmerket*, and *utsøkt*.

We have also decided to omit all multi-word expressions, and only keep single-word translations. For example the translations of the negative-labeled expressions *die-hard*, *layoff-happy*, *ultra-hardline*, *muscle-flexing*, *martyrdom-seeking*, *anti-israeli*; and positive-labeled expressions like *counter-attacks*, *well-positioned*, and *well-backlit* were not included.

Some other words were not translated because we either believed that they did not fit into the positive or negative categories, or because we could not find an appropriate Norwegian translation. Examples of some of the originally negative-labeled words that fell into these categories are: *drones*, *vibration*, *miscellaneous*, *frost*, *funny*, *flirt*, *sober*, and *rhetorical*. Examples of positive-labeled words that were excluded are *work*, *hotcakes*, *rapport*, *dawn*, *illuminati*, *electrify*, *ftw*, and *instrumental*. We also removed all words that were present in both the positive and the negative lists. This process resulted in a Norwegian sentiment lexicon containing a collection of 3,917 negative and 1,601 positive words. Table 1 gives an overview of the word classes present in the translated Norwegian lexicon (*Translated*). Several words can overlap between word classes, for example 60 positive nouns and 123 negative nouns are also adjectives.

| | | Negative | Positive |
|---|---|---|---|
| *Translated* | All | 3,917 | 1,601 |
| | Adjectives | 1,728 | 844 |
| | Verbs | 1,575 | 541 |
| | Nouns | 1,371 | 461 |
| | Participle adjectives | 146 | 97 |
| *Full-forms* | All | 14,839 | 6,103 |
| | Adjectives | 6,392 | 3,030 |
| | Verbs | 5,769 | 2,269 |
| | Nouns | 4,565 | 1,559 |
| | Participle adjectives | 938 | 368 |
| *Lemmas* | All | 4,939 | 2,004 |
| | Adjectives | 2,085 | 958 |
| | Verb | 942 | 371 |
| | Noun | 1,186 | 415 |
| | Participle adjectives | 934 | 366 |

Table 1: Overview of the Norwegian sentiment lexicon, showing counts for the manually inspected translations, the full-forms of the expanded version, and finally the lemmas found after expansion.

Similarly to the English lexicon, the resulting Norwegian lexicon contains a mix of word classes and inflected forms. In order to produce a more general version of the lexicon containing all possible word-forms (*Full-forms*), we have used the previously mentioned Norwegian full-form lexicon SCARRIE to expand the entries to include all inflected forms. This resulted in a lexicon of 14,839 negative words and 6,103 positive words. Table 1 gives a detailed overview of the content of the Norwegian lexicon, both with regards to the

number of word-forms and lemmas, where participle adjectives are all words that can be both adjectives and participles.

Our preliminary experiments showed that using the Norwegian lexicon as directly translated yields similar results to using the expanded lexicon. In what follows we therefore only report results of using the translated and manually curated (but non-expanded) Norwegian lexicon. However, we make both versions of the lexicon publicly available.[5]

Additionally, we set aside 20 percent of each lexicon (1,357 words for English, 1,122 for Norwegian) as a development set to monitor the performance of the models on the auxiliary task.

## 4 Sentiment datasets

In the following we present the datasets used to train and evaluate our sentence-level classifiers.

### 4.1 English

The Stanford Sentiment Treebank (SST) (Socher et al., 2013) contains 11,855 sentences taken from English-language movie reviews. It was annotated for fine-grained sentiments (strong negative, negative, neutral, positive, strong positive) based on crowdsourcing. We perform experiments using the pre-defined train, development and test splits (of 8,455 / 1,101 / 2,210 sentences, respectively).

### 4.2 Norwegian

The Norwegian dataset used in this work forms part of the Norwegian Review Corpus NoReC (Velldal et al., 2018), consisting of full-text reviews from a range of different domains, such as restaurants, literature, and music, collected from several of the major Norwegian news sources. The particular subset used in the current work, dubbed NoReC$_{eval}$, comprises 7961 sentences across 298 documents that have been manually annotated according to whether or not each sentence contains an *evaluation*, as described by Mæhlum et al. (2019). Two types of evaluative sentence categories are distinguished (in addition to non-evaluative sentences): simple *evaluative* and a special case of *evaluative fact-implied non-personal*. The latter follows the terminology of Liu (2015), denoting factual, objective sentences which are used with an evaluative intent but without reference to personal experience. Example (1) shows an evaluative sentence, labeled `EVAL`, which contains the positive evaluation signaled by the adjectives *sterk* 'strong/powerful' and *flott* 'great'.

(1) *Sterk og flott film om hevntanker*
Strong and great movie about revenge
A powerful and great movie about revenge

Example (2) shows a fact-implied non-personal sentence, labeled `FACT-NP`, where a factual, objective statement is interpreted as expressing an evaluation given the context of a car review.

(2) *Firehjulsdriften kan kobles inn og ut etter behov.*
Fourwheeldrive can switched in and out after need
The four wheel drive can be switched on and off as required

Unlike the English dataset discussed above, the annotation does not specify the polarity of the sentence. The rationale for this is that a sentence may contain more than one sentiment expression and have a mixed polarity, hence this type of annotation is better performed sub-sententially following an initial annotation of evaluative or sentiment-relevant sentences (Toprak et al., 2010; Scheible and Schütze, 2013).

We use the training, development and test splits as defined by Mæhlum et al. (2019), see the summary of corpus statistics in Table 2.

## 5 Multi-task learning of lexicon information in neural models

This section details our multi-task neural architecture for incorporating sentiment lexicon information into neural networks, as shown in Figure 1. Our multi-task model shares the lower layers (an embedding and fully connected layer), while allowing the higher layers to further adapt to the main and auxiliary tasks. Specifically, we use a sentiment prediction auxiliary task, where the goal is to correctly predict whether a single word is positive or negative, to improve the main task of sentence-level classification. Although the units of classification for the two tasks are different (word-level in the auxiliary task and sentence-level in the main), the auxiliary task can be assumed to be highly predictive for the main task, as sentiment bearing words are the main feature for identifying evaluative sentences and their polarity.

[5] https://github.com/ltgoslo/norsentlex

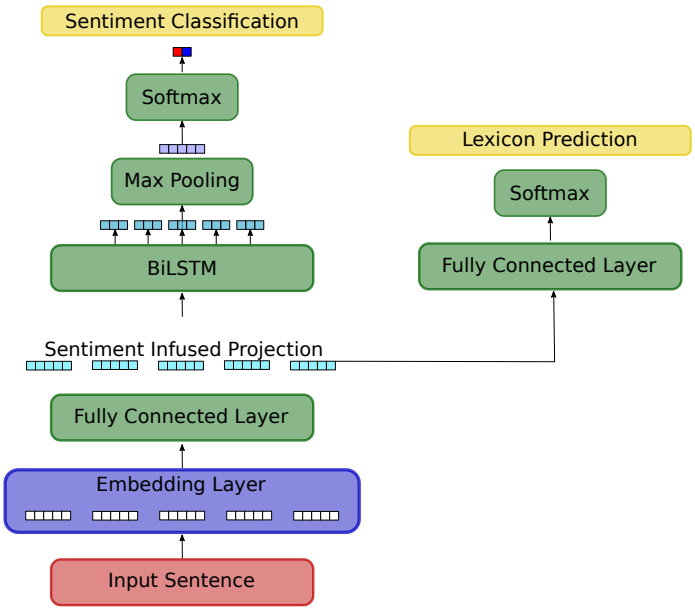


Figure 1: Our proposed multi-task model to incorporate lexicon information into a neural classifier.

| | Train | Dev | Test |
|---|---|---|---|
| Documents | 230 | 33 | 35 |
| Sentences | 5,915 | 1,151 | 895 |
| Tokens | 104,350 | 20,755 | 16,292 |

Table 2: Corpus counts for the Norwegian dataset.

**Lexicon prediction model:** We propose a lexicon prediction model, which given a word from a sentiment lexicon, predicts whether the word is positive or negative. We implement a multi-layer feed-forward network which uses word embeddings as input, ReLU non-linearities, and a softmax layer for classification. This model has previously shown promise for predicting abstractness (Köper and Schulte im Walde, 2017) and emotion ratings (Köper et al., 2017). We additionally use dropout (0.3) after the embedding layer for regularization.

**Sentence-level prediction model:** For sentiment classification, we use a bidirectional Long Short-Term Memory network to create contextualized representations of each token after being projected to the sentiment infused space. The final contextualized vectors at each time step are concatenated and then passed to a max pooling layer and finally to a softmax layer for classification. This single-task model trained without the lexicon prediction task (STL) is also used as a baseline to measure the relative improvement.

**Multi-task model:** During training, the multi-task learning model (MTL) alternates between training one epoch on the main task and one epoch on the auxiliary task. Preliminary experiments showed that more complicated training strategies (alternating training between each batch or uniformly sampling batches from the two tasks) did not lead to improvements. For English we use 300 dimensional pre-trained embeddings from GoogleNews,[6] while for Norwegian we use 100 dimensional skip-gram fastText embeddings (Bojanowski et al., 2016) trained on the NoWaC corpus (Guevara, 2010). The pre-trained embeddings were re-used from the NLPL vector repository[7] (Fares et al., 2017). We train the model for 10 epochs using Adam (Kingma and Ba, 2014), performing early stopping determined by the improvement on the development set of the main task. Given that neural models are sensitive to the random initialization of their parameters, we perform five runs with different random seeds and show the mean and standard deviation as the final result for each model. We use the same five random seeds for all experiments to ensure a fair comparison between models.

[6]Available at https://code.google.com/archive/p/word2vec/.

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

| Model | SST | NoReC$_{eval}$ |
|---|---|---|
| LEXICON | 14.7 | 37.2 |
| BOW | 37.4 | 45.0 |
| BOW+LEXICON | 38.9 | 45.8 |
| LEX-EMB | 34.7 (1.1) | 48.9 (0.1) |
| STL | 37.8 (3.1) | 51.2 (2.6) |
| MTL | **42.4 (3.2)** | **52.8 (2.9)** |

Table 3: Macro $F_1$ of models on the SST and NoReC$_{eval}$ sentence-level datasets. Neural models report mean and standard deviation of the scores over five runs.

**Lexicon embedding Model:** We also include an additional model (LEX-EMB), which uses the feed-forward network previously described to learn to predict word-level polarity and the same BiLSTM architectures for sentiment classification. Instead of jointly learning the two tasks, we first train the feed-forward model on the lexicon task and update the original embeddings. We then concatenate these learned sentiment embeddings to the original embeddings to create a sentiment informed representation of each word before passing them to the BiLSTM. All other parts of the models are the same as the STL model.

**Baselines:** We also include three non-neural baseline models: LEXICON, BOW, BOW+LEXICON. The LEXICON baseline uses the sentiment lexicon to create features for a Linear SVM. The inputs to the SVM were presented as sequences of labels representing the sentences of the datasets, such that each word present in the lexicons was labeled as either $+1$ or $-1$ for positive or negative words respectively, and the rest were labeled as $0$. This was done to incorporate lexicon information, and predict classes based on the distribution of positive and negative words within sentences. The BOW baseline uses a bag-of-words representation of the data to train a Linear SVM. Finally, the BOW+LEXICON adds two additional features to the BOW model: the total number of tokens which are found in the positive and negative lists in the sentiment lexicons. We choose the optimal $c$ value for each classifier on the development split.

## 6 Model results

Table 3 shows the Macro $F_1$ scores for all models on the SST and NoReC$_{eval}$ test sets. Note that previous work often uses accuracy as a metric on the SST dataset, but as we hypothesize that lexicon information may help the minority classes, we thought it important to give these equal weight. (A macro $F_1$ of 40.1 in our case corresponds to 46.7 accuracy). The BOW model performs quite well on SST, only 0.4 percentage points (ppt) worse than STL. On NoReC$_{eval}$, however, it performs much worse, which can be attributed to the the difficulty of determining if a sentence is non-evaluative or fact-implied using only unigram information, as these sentence types do not differ largely lexically.

BOW+LEXICON performs better than BOW on both datasets, although the difference is larger on SST (1.5 ppt vs. 0.8 ppt). This is likely due to sentiment lexicon features being more predictive for the sentiment task. Additionally, it outperforms the STL model by 1.1 ppt on SST, confirming that it is a strong baseline (Mohammad et al., 2013).

LEX-EMB is the weakest model on the SST dataset with 34.7 $F_1$ but performs better than the non-neural baselines on NoReC$_{eval}$ (48.9). STL performs better than LEXICON, BOW, and LEX-EMB on both tasks, as well as BOW+LEXICON on NoReC$_{eval}$. Finally, MTL is the best performing model on both tasks, with a difference of 3.5 ppt between MTL and the next best performing model on SST, and 1.6 ppt on NoReC$_{eval}$.

### 6.1 Error analysis

We perform an error analysis by comparing how the MTL model changes predictions when compared to the STL model. We create a confusion matrix of the predictions of each model on the SST and NoReC$_{eval}$ tasks over all five runs and show the relative differences in Figures 2 and 3. Positive numbers (dark purple) indicate that the MTL model made more predictions in this square, while negative numbers (white) indicate it made fewer predictions.

Counter-intuitively, the MTL model improves mainly on the neutral, strong positive, and strong negative classes, while performing relatively worse on the positive and negative classes. In general, the MTL makes fewer negative and positive predictions than the STL model. On the NoReC$_{eval}$ task, the MTL model leads to fewer ab-

| Sentence | Gold | STL | MTL |
|---|---|---|---|
| Light, cute and forgettable. | neutral | negative | neutral |
| Despite some gulps the film is a fuzzy huggy. | positive | negative | positive |
| This is art paying homage to art. | positive | positive | neutral |
| Undercover Brother doesn't go far enough. | negative | negative | neutral |

Table 4: Examples where MTL performs better and worse than STL. A red box indicates negative polarity (blue box indicates positive) according to the sentiment lexicon used to in the auxiliary training task.

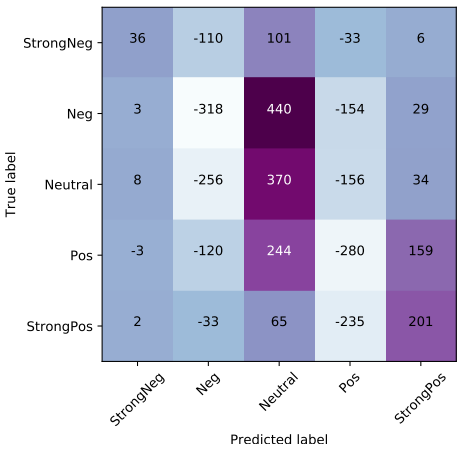


Figure 2: A relative confusion matrix on the SST task, where positive numbers (dark purple) indicate that the MTL model made more predictions than STL in the square and negative (white) indicate that it made fewer.

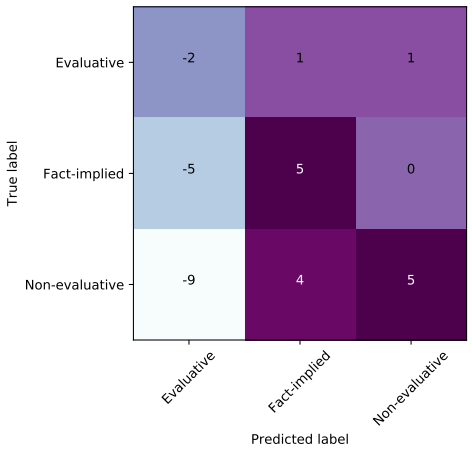


Figure 3: A relative confusion matrix on the NoReC$_{eval}$ task, where positive numbers (dark purple) indicate that the MTL model made more predictions than STL in the square and negative (white) indicate that it made fewer.

| Model | English | Norwegian |
|---|---|---|
| LEX-EMB | **87.1 (0.2)** | **87.5 (0.2)** |
| MTL | 86.7 (0.2) | 72.4 (3.9) |

Table 5: Mean accuracy and standard deviation of the MTL and LEX-EMB models over five runs on the Hu and Liu lexicon for English and the translated lexicon for Norwegian.

solute changes, but importantly reduces the number of non-evaluative sentences predicted as evaluative. Again, the MTL model has a tendency to reduce predictions for the majority class and increase them for the minority classes (fact-implied and non-evaluative). This seems to point to the regularizing effect of multi-task learning (Augenstein et al., 2018; Bingel and Søgaard, 2017). Table 4 additionally shows examples where MTL is better and worse than STL.

### 6.2 Lexicon prediction results

In this section, we evaluate the performance of the MTL and LEX-EMB models on the auxiliary lexicon prediction task. Table 5 shows that the LEX-EMB model outperform the MTL model on both English and Norwegian. For English the difference between models is small (0.4 ppt), while much larger for Norwegian (15.1 ppt). Rather than being attributed to differences in language, we hypothesize that the difference is due to task similarity. For English, the auxiliary task is much more predictive of the main task (sentence-level sentiment), while for Norwegian the main task of predicting evaluative, fact implied and non-evaluative does not depend as much on word-level sentiment. The MTL classifier in Norwegian therefore relies less on the auxiliary module.

| Model | Lexicon | # tokens | SST |
|---|---|---|---|
| STL | – | – | 37.8 (3.1) |
| MTL | SOCAL | 4,539 | 41.3 (3.1) |
| | SOCAL GOOGLE | 1,691 | 37.9 (0.3) |
| | NRC EMOTION | 4,460 | 41.5 (3.1) |
| | HU AND LIU | 5,432 | **42.4 (3.2)** |

Table 6: Macro $F_1$ of models on the SST sentence-level datasets. We compare the MTL model on SST using different lexicons.

### 6.3 Other lexicons

In this section, we experiment with using different English lexicons as an auxiliary task for the MTL model. Specifically, we compare the following sentiment lexicons:

- SOCAL: a sentiment lexicon compiled manually with words taken from several review domains (Taboada et al., 2011).
- SOCAL GOOGLE: a semi-supervised lexicon created from a small set of seed words (Taboada et al., 2006) using a PMI-based technique and search engine queries (Turney and Littman, 2003).
- NRC EMOTION: a crowd-sourced emotion lexicon which also contains polarity annotations (Mohammad and Turney, 2013).
- HU AND LIU: the sentiment lexicon described in § 2.

While the NRC EMOTION lexicon already contains binary annotations, tokens in SOCAL and SOCAL GOOGLE are annotated on a scale from $-5$ to $5$. We make these annotations binary by assigning positive polarity to tokens with a rating $> 0$ and negative for those $< 0$. Any neutral tokens are discarded. Table 6 shows that the MTL model is robust to different sources of sentiment information. The size of the dataset appears to be more important than the specific content, as all lexicons over 4,000 words achieve similar scores.

## 7 Conclusion

This paper proposes a method to incorporate external knowledge, in this case about word polarity in the form of sentiment lexicons, into a neural classifier through multi-task learning. We have performed experiments on sentence-level sentiment tasks for English and Norwegian, demonstrating that our multi-task model improves over a single-task approach in both languages. We provide a detailed analysis of the results, concluding that the multi-task objective tends to help the neutral and minority classes, indicating a regularizing effect.

We have also introduced a Norwegian sentiment lexicon, created by first machine-translating an English lexicon and manually curating the results. This lexicon, and its expansion to a full-form lexicon, are made freely available to the community. While our current model ignores subword information, *e.g. unimpressive*, and multiword expressions, *e.g. not my cup of tea*, including this information could further improve the results.

Although we have limited the scope of our auxiliary task to binary classification, using a regression task with sentiment and emotion labels may provide more fine-grained signal to the classifier. We also plan to experiment with a similar setup for targeted or aspect-level classification tasks.

Finally, it is important to note that the MTL approach outlined in this paper could also be applied to incorporate other types of external knowledge into neural classifiers for other types of tasks besides sentiment analysis.

## 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).

## References

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. 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 1896–1906, New Orleans, Louisiana.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-Task Learning of Keyphrase Boundary Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 341–346, Vancouver, Canada.

Aleksander Bai, Hugo Hammer, Anis Yazidi, and Paal Engelstad. 2014. Constructing sentiment lexicons in Norwegian from a large text corpus. In *Proceedings of the 17th IEEE International Conference on Computational Science and Engineering*, pages 231–237, Chengdu, China.

Lingxian Bao, Patrik Lambert, and Toni Badia. 2019a. 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. Association for Computational Linguistics.

Lingxian Bao, Patrik Lambert, and Toni Badia. 2019b. 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.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain.

Johannes Bjerva. 2017. Will my auxiliary tagging task help? Estimating Auxiliary Tasks Effectivity in Multi-Task Learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220, Gothenburg, Sweden.

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

Margaret M. Bradley, Peter J. Lang, Margaret M. Bradley, and Peter J. Lang. 1999. Affective norms for English Words (ANEW): Instruction manual and affective ratings.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, volume 6, pages 417–422, Genova, Italy.

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.

Murhaf Fares, Stephan Oepen, and Erik Velldal. 2018. Transfer and multi-task learning for nounnoun compound interpretation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498, Brussels, Belgium.

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.

Hugo Hammer, Aleksander Bai, Anis Yazidi, and Paal Engelstad. 2014. Building sentiment lexicons applying graph theory on information from three Norwegian thesauruses. In *Norsk Informatikkonferanse (NIK)*, Fredrikstad, Norway.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, USA.

Ozan Irsoy and Claire Cardie. 2014. Deep Recursive Neural Networks for Compositionality in Language. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2096–2104. Curran Associates, Inc.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3190–3195, Istanbul, Turkey.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain.

Zeyang Lei, Yujiu Yang, and Min Yang. 2018a. Sentiment lexicon enhanced attention-based lstm for sentiment classification. In *AAAI-2018-short paper*, page 81058106.

Zeyang Lei, Yujiu Yang, Min Yang, and Yi Liu. 2018b. A multi-sentiment-resource enhanced attention network for sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 758–763, Melbourne, Australia.

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

Petter Mæhlum, Jeremy Claude Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Turku, Finland.

Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. 2019. Attention-based conditioning methods for external knowledge integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3944–3951, Florence, Italy.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, USA.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages*, pages 93–98, Heraklion, Crete.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, Michigan.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada.

Barbara Plank. 2016. Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 609–619, Osaka, Japan.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden.

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.

Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–158, Copenhagen, Denmark.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 231–235, Berlin, Germany.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for Creating Semantic Orientation Dictionaries. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Zhiyang Teng, Duy-Tin Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1629–1638, Austin, Texas. Association for Computational Linguistics.

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.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Erik Velldal, Lilja Øvrelid, Cathrine Stadsnes Eivind Alexander Bergem, Samia Touileb, and Fredrik Jrgensen. 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, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada.

Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2018. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 868–877, Santa Fe, New Mexico, USA. Association for Computational Linguistics.