Multi-task Learning of Negation and Speculation for Targeted Sentiment Classification

Andrew Moore*  
School of Computing and Communications, Lancaster University  
a.moore@lancaster.ac.uk

Jeremy Barnes  
University of Oslo  
Department of Informatics  
jeremycb@ifi.uio.no

Abstract

The majority of work in targeted sentiment analysis has concentrated on finding better methods to improve the overall results. Within this paper we show that these models are not robust to linguistic phenomena, specifically negation and speculation. In this paper, we propose a multi-task learning method to incorporate information from syntactic and semantic auxiliary tasks, including negation and speculation scope detection, to create models that are more robust to these phenomena. Further we create two challenge datasets to evaluate model performance on negated and speculative samples. We find that multi-task models and transfer learning from a language model can improve performance on these challenge datasets. However the results indicate that there is still much room for improvement in making our models more robust to linguistic phenomena such as negation and speculation.

1 Introduction

Targeted sentiment analysis (TSA) involves jointly predicting entities which are the targets of an opinion, as well as the polarity expressed towards them (Mitchell et al., 2013). The TSA task, which is part of the larger set of fine-grained sentiment analysis tasks, can enable companies to provide better recommendations (Bauman et al., 2017), as well as give digital humanities scholars a quantitative approach to identifying how sentiment and emotions develop in literature (Alm et al., 2005; Kim and Klinger, 2019).

Although there have been many improvements to modelling TSA since the original CRF models (Mitchell et al., 2013), such as utilising Recurrent Neural Networks (RNN) (Zhang et al., 2015a; Katiyar and Cardie, 2016; Ma et al., 2018), and treating the task as span prediction rather than a sequence labelling task (Hu et al., 2019), most of these have concentrated on making the best use of data annotated specifically for the task. However, annotation for fine-grained sentiment is more taxing and tends to have lower inter-annotator agreement than document or sentence classification tasks (Wiebe et al., 2005; Øvrelid et al., 2020). This leads to a lack of available high-quality training data, even for highly resourced languages and prevents TSA models from learning the complex, compositional phenomena which are necessary to correctly predict targeted sentiment in an end-to-end fashion.

We believe this lack of data for fine-grained sentiment analysis leads to TSA models that cannot learn effectively complex compositional phenomena that exists in language, thus making TSA models fragile to highly compositional language. It has also been shown that incorporating compositional information from negation or speculation detection improves sentence-level sentiment classification (Councill et al., 2010; Cruz et al., 2016; Barnes et al., 2020). Other supervised tasks, such as semantic role labelling (Marasović and Frank, 2018), or document level sentiment analysis (He et al., 2019) have shown promise for improving fine-grained sentiment analysis. Further transfer learning from a self-supervised language-modelling task, commonly referred to as contextualised word representations (CWR), has also shown to greatly benefit fine-grained sentiment analysis (Hu et al., 2019; Li et al., 2019b). Based on this, in this paper, we wish to explore two research questions:

1. Does multi-task learning of negation and speculation lead to more robust targeted sentiment models?

*The authors contributed equally.
2. Does transfer learning based on language-modelling already incorporate this information in a way that is useful for targeted sentiment models?

To this end, we propose a multi-task learning (MTL) approach to incorporate sources of negation and speculation information into a neural targeted sentiment classifier. We additionally compare our approach with MTL models that use part-of-speech tagging, dependency relation prediction, and lexical analysis as auxiliary tasks, following previous work (Wang and Pan, 2018a). Furthermore, in order to overcome the lack of evaluative resources to investigate the effects of negation and speculation, we annotate two new challenge datasets which contain difficult negated and speculative examples.

We find that the MTL models are more robust than the single task learning (STL), performing competitively on the majority of the standard datasets while significantly outperforming the STL models on the negation challenge datasets, and on average better than STL models on the speculation challenge datasets. Moreover, we show that when transfer learning is applied, using CWR, to both MTL and STL models, MTL models are no longer significantly better, but are still better on average for the negation challenge dataset and one of the speculation challenge datasets. This result suggests that transfer learning does incorporate some compositional information that is required for negated and speculative samples. However all results on the challenge datasets are considerably lower than the standard dataset, showing that more work is needed to make these models more robust to compositional language.

The contributions of the paper are the following: i) We introduce two English challenge datasets that are annotated for negation and speculation. ii) We propose a multi-task model to incorporate negation and speculation information and evaluate it across four English datasets. iii) Finally, using the challenge dataset, we are the first to show the quantitative effect of negation and speculation on TSA.

2 Background and related work

Fine-grained sentiment analysis is a complex task which can be broken into four subtasks (Liu, 2015): i) opinion holder extraction, ii) opinion target extraction, iii) opinion expression extraction, iv) and resolving the polarity relationship between the holder, target, and expression. From these four subtasks, TSA (Jin and Ho, 2009; Chen et al., 2012; Mitchell et al., 2013) reduces the fine-grained task to only the second and final subtasks ignoring the extraction of holder and sentiment expressions. The TSA task therefore involves identifying sentiment targets and simultaneously predicting their polarity, given a context – normally the full text of a sentence.

There are numerous datasets that can be used for TSA, English datasets including MPQA (Wiebe et al., 2005), the SemEval Laptop and Restaurant reviews (Pontiki et al., 2014; Pontiki et al., 2016), and Twitter datasets (Mitchell et al., 2013; Wang et al., 2017). Further annotation projects have led to review datasets for Arabic, Dutch, French, Russian, and Spanish (Pontiki et al., 2016) and Twitter datasets for Spanish (Mitchell et al., 2013) and Turkish (Pontiki et al., 2016). Prior work has also explored the effects of different phenomena on TSA through error analysis and challenge datasets. Wang et al. (2017), Xue and Li (2018), and Jiang et al. (2019) showed the difficulties of polarity classification of targets on texts with multiple different polarities through the distinct sentiment error splits, the hard split, and the MAMS challenge dataset respectively. Both Kaushik et al. (2020) and Gardner et al. (2020) also augment document sentiment datasets by asking annotators to make changes that would lead to counterfactual examples for IMDB dataset. Most recently Ribeiro et al. (2020) showed how sentence level sentiment models are affected by various linguistic phenomena including negation, semantic role labelling, temporal changes, and name entity recognition. However, none of these works have quantified the effect of negation and speculation on TSA models.

Previous approaches to modelling TSA have often relied on general sequence labelling models, e.g. Conditional Random Fields (CRF) (Mitchell et al., 2013), probabilistic graphical models (Klinger and Cimiano, 2013), RNN (Zhang et al., 2015b; Ma et al., 2018), and more recently self-attention networks pretrained on language modelling (Li et al., 2019b). Additionally Li et al. (2019a) used two auxiliary tasks, target extraction and an opinion word based task, to guide the main TSA task. However none of the previous works within TSA has consider negation or speculation as additional information for a model nor considered quantitatively evaluating the effect it could have.
Multi-task and transfer learning  The main idea of MTL (Caruana, 1993) is that a model which receives signal from two or more correlated tasks will more quickly develop a useful inductive bias, allowing it to generalize better. This approach has gained traction in NLP, where several benchmark datasets have been created (Wang et al., 2019b; Wang et al., 2019a). Under some circumstances, MTL can also be seen as a kind of data augmentation, where a model takes advantage of extra training data available in an auxiliary task to improve the main task (Kshirsagar et al., 2015; Plank, 2016).

Much of MTL uses hard parameter sharing (Caruana, 1993), which shares all parameters across some layers of a neural network. When the main task and auxiliary task are closely related, this approach has been shown to be an effective way to improve model performance (Collobert et al., 2011; Peng and Dredze, 2017; Martinez Alonso and Plank, 2017; Augenstein et al., 2018), although it is often preferable to make predictions for low-level auxiliary tasks at lower layers of a multi-layer MTL setup (Søgaard and Goldberg, 2016), which we refer to as hierarchical MTL. More recently, soft parameter sharing (Liu et al., 2016; Ruder et al., 2019) and multi-task knowledge distillation (Ba and Caruana, 2013; Hinton et al., 2014; Clark et al., 2019) have shown further improvements by relaxing the assumption that all layers share parameters. However, these approaches do not often lead to large improvements over hierarchical MTL (Marasović and Frank, 2018; Ruder et al., 2019).

Specific to TSA, previous research has used MTL to incorporate others sources of information into their models, e.g. document level sentiment (He et al., 2019), dependency parsing for cross domain (Wang and Pan, 2018a) and cross lingual (Wang and Pan, 2018b) target extraction, and semantic role labelling for target and holder extraction (Marasović and Frank, 2018; Zhang et al., 2019). However none of these approaches have investigated whether negation and speculation are useful auxiliary tasks for TSA. Additionally, transfer learning methods (Mikolov et al., 2013; Peters et al., 2018a; Devlin et al., 2019) can leverage unlabeled data from large monolingual corpora, but minimally require training large models on unlabeled corpora. However, it is not clear to what extent these models implicitly learn negation and speculation composition through the language modelling objective.

Negation and Speculation Detection  As negation is such a common linguistic phenomenon and one that has a direct impact on sentiment, previous work has shown that incorporating negation information is crucial for accurate sentiment prediction. Feature-based approaches did this by including features from negation detection modules (Das and Chen, 2007; Councill et al., 2010; Lapponi et al., 2012), while it has now become more common to assume that neural models learn negation features in an end-to-end fashion (Socher et al., 2013). However, recent research suggests that end-to-end models are not able to robustly interpret the effect of negation on sentiment (Barnes et al., 2019), and that explicitly learning negation can improve sentiment results (Barnes, 2019; Barnes et al., 2020).

On the other hand, speculation refers to whether a statement is described as a fact, a possibility, or a counterfact (Saurí and Pustejovsky, 2009). Although there are fewer speculation annotated corpora available (Vincze et al., 2008; Kim et al., 2013; Konstantinova et al., 2012), including speculation information has shown promise for improving sentiment analysis at document-level (Cruz et al., 2016).

There has, however, been little research on how these phenomena specifically affect fine-grained approaches to sentiment analysis. This is important because, compared to document- or sentence-level tasks where there is often a certain redundancy in sentiment signal, for fine-grained tasks negation and speculation often completely change the sentiment (see Table 2), making their identification and integration within a fine-grained sentiment models essential to resolve.

3 Data

We perform the main experiments on four English language datasets: The Laptop dataset from SemEval 2014 (Pontiki et al., 2014), the Restaurant dataset which combines the SemEval 2014 (Pontiki et al., 2014), 2015 (Pontiki et al., 2015), and 2016 (Pontiki et al., 2016), the Multi-aspect Multi-sentiment (MAMS) dataset (Jiang et al., 2019), and finally the Multi-perspective Question Answering (MPQA) dataset (Wiebe et al., 2005). We take the pre-processed Laptop and Restaurant datasets from Li et al.\(^1\)

\(^1\)All datasets contain the following three sentiment classes positive, neutral, and negative. The MPQA dataset also includes a fourth rare class, both. The distribution of the sentiment classes can be seen in table 7 of Appendix A.
Train | Dev | Test
--- | --- | ---
\(sents.\) | \(targs.\) | \(len.\) | \(mult.\) | \(sents.\) | \(targs.\) | \(len.\) | \(mult.\) | \(sents.\) | \(targs.\) | \(len.\) | \(mult.\)
Laptop | 2,741 | 2,044 | 1.5 | 136 | 304 | 256 | 1.5 | 18 | 800 | 634 | 1.6 | 38
Laptop\(_{Neg}\) | – | – | – | – | 147 | 181 | 1.5 | 41 | 403 | 470 | 1.6 | 79
Laptop\(_{Spec}\) | – | – | – | – | 110 | 142 | 1.4 | 10 | 208 | 220 | 1.5 | 19
Restaurant | 3,490 | 3,896 | 1.4 | 312 | 387 | 414 | 1.4 | 34 | 2,158 | 2,288 | 1.4 | 136
Restaurant\(_{Neg}\) | – | – | – | – | 198 | 274 | 1.4 | 61 | 818 | 1,013 | 1.4 | 161
Restaurant\(_{Spec}\) | – | – | – | – | 138 | 200 | 1.3 | 35 | 400 | 451 | 1.4 | 49
MAMS | 4,297 | 11,162 | 1.3 | 4,287 | 500 | 1,329 | 1.3 | 498 | 500 | 1,332 | 1.3 | 500
MPQA | 4,195 | 1,264 | 6.3 | 94 | 1,389 | 400 | 5.4 | 29 | 1,620 | 365 | 6.7 | 22

Table 1: Statistics for the sentiment datasets used in the experiments. The table indicates the number of sentences in each split (\(sents.\)), the number of targets (\(targs.\)), the average length of the targets (\(len.\)), as well as how many sentences in each have multiple targets with differing polarity (\(mult.\)).

| original | this is good, inexpensive sushi. | negated | this is not good, inexpensive sushi. | speculative | I’m not sure if this is good, inexpensive sushi. |
| --- | --- | --- | --- | --- | --- |

Table 2: Example of how adding negation and speculation can change the polarity of a target (added tokens are shown in bold). While in the original, the target “sushi” has a positive polarity, in the negated example it is negative, and in the speculative example it is neutral.

and use the train, dev, and test splits that they provide. The NLTK word tokenizer was used to tokenise the Laptop, Restaurant, and MPQA datasets whereas Spacy was used for the MAMS dataset.

We choose datasets that differ largely in their domain, size, and annotation style in order to determine if any trends we see are robust to these data characteristics or whether they are instead correlated. While the Laptop, Restaurant and MAMS datasets were annotated for aspect-based sentiment analysis, it is simple to convert them to a TSA setup by extracting the aspect targets and their polarity. We do the same for the MPQA dataset and convert all data to BIOUL format with unified sentiment tags, e.g. B-POS for a beginning tag with a positive sentiment, so that we can cast the TSA problem as a sequence labeling task. We use the unified tagging scheme following recent work (Li et al., 2019a; Li et al., 2019b).

The statistics for these datasets are shown in Table 1. MAMS has the largest number of training targets (11,162), followed by Restaurant (3,896), Laptop (2,044) and finally MPQA has the fewest (1,264). MPQA, however, has the longest average targets (6.3 tokens) compared to 1.3-1.5 for the other datasets. This derives from the fact that entire phrases are often targets in MPQA. Finally, due to the annotation criteria, the MAMS data also has the highest number of sentences with multiple aspects with multiple polarities – nearly 100% in train, compared to less than 10% for Restaurant.

3.1 Annotation for negation and speculation

Although negation and speculation are prevalent in the original data – negation and speculation occur in 13-25% and 9-20% of the sentences, respectively – it is difficult to pry apart improvement on the dev and test data with improvement on these two phenomena. Therefore, we further annotate the dev and test set for the Laptop and Restaurant dataset\(^3\), and when possible, insert negation and speculation cues into the sentence, which we call Laptop\(_{Neg}\), Laptop\(_{Spec}\), Restaurant\(_{Neg}\), and Restaurant\(_{Spec}\). Inserting negation and speculation cues often leads to a change in polarity from the original annotation, as shown in the example in Table 2. We finally keep all sentences that contain a negation or speculation cue, respectively, including those that occur naturally in the data. The statistics are shown in Table 1\(^4\).

\(^2\)This is also known as collapsed tagging scheme (Hu et al., 2019)

\(^3\)For clarification this is the SemEval 2014 Laptop dataset and the 2014, 2015, and 2016 combined Restaurant dataset.

\(^4\)The distribution of the sentiment classes can be seen in table 7 of Appendix A.
### 3.2 Auxiliary task data

For the multi-task learning experiments, we use six auxiliary tasks: negation scope detection using the Conan Doyle (CD) \cite{Morante2012}, both negation detection (SFU) and speculation detection (SPEC) on the SFU\textsubscript{NegSpec} dataset \cite{Konstantinova2012}, and Universal Part-of-Speech tagging (UPOS), Dependency Relation prediction (DR) and prediction of full lexical analysis (LEX) on the Streusle dataset \cite{Schneider2015}. We show the train, dev, test splits, as well as the number of labels, label entropy and label kurtosis \cite{MartinezAlonso2017} in Table 3. An example sentence with auxiliary labels is shown in Appendix B. Although it may appear that the SFU dataset is an order of magnitude larger than the Conan Doyle dataset, in reality, most of the training sentences do not contain annotations, leaving similar sized data if these are filtered. Similar to the sentiment data, we convert the auxiliary tasks to BIO format and treat them as sequence labelling tasks.

### 4 Experiments

We experiment with a single task baseline (STL) and a hierarchical multi-task model with a skip-connection (MTL), both of which can be seen in Figure 1. For the STL model, we first embed a sentence and then pass the embeddings to a Bidirectional LSTM (Bi-LSTM). These features are then concatenated to the input embeddings and fed to the second Bi-LSTM layer, ending with the token-wise sentiment predictions from the CRF tagger. For the MTL model, we additionally use the output of the first Bi-LSTM layer as features for the separate auxiliary task CRF tagger. As can be seen from Figure 1, the STL model and the MTL main task model use the same the green layers. The MTL additionally uses the pink layer for the auxiliary task, adding less than 3.4% trainable parameters\footnote{The STL model had 1,785,967 parameters of which 364,042 were trainable as the embedding layer was frozen.} for all auxiliary tasks except LEX, which adds 221.4% due to the large label set (see Table 3). Furthermore, at inference time the MTL model is as efficient as STL, given that it only uses the green layers when predicting the targeted sentiment, of which this is empirically shown in Table 20 of Appendix F.

**Embeddings:** For the embedding layer, we perform experiments using 300 dimensional GloVe embeddings \cite{Pennington2014}, as well as CWR from Transformer ELMo embeddings \cite{Peters2018}. The GloVe embeddings are publicly available and trained on the concatenation of English Wikipedia and Gigaword data. For the MPQA dataset we used the standard Transformer ELMo from Peters et al. \cite{Peters2018} which was trained on the 1 billion word benchmark \cite{Chelba2014}. For experiments on the MAMS and Restaurant datasets we further tuned the standard Transformer ELMo on 27 million (M) sentences from the 2019 Yelp review dataset\footnote{This can be found here \url{https://www.yelp.com/dataset}} and for the Laptop dataset on 28M sentences\footnote{More specifically there was 9M unique sentences and the model was trained for 3 epochs.} from the Amazon electronics reviews dataset \cite{McAuley2015}. For all experiments...
Figure 1: The overall architecture where the STL model contains all of the green layers and the MTL uses the additional pink auxiliary CRF tagger. The second Bi-LSTM has a skip connection from the embedding layer which concatenates the word embeddings with the output from the first Bi-LSTM.

we freeze the embedding layer in order to make the results between GloVe and CWR more comparable with respect to the number of trainable parameters. For the CWR instead of using the output from the last layer, we learn a summed weighting of all layers\(^1\), as this has been shown more effective than using the last layer (Peters et al., 2018a). For more details on the number of parameters used for each model see Table 19 in Appendix F.

Training: For the STL and the MTL models, we tune hyperparameters using AllenTune (Dodge et al., 2019) on the Laptop development dataset. We then use the best hyperparameters on the the Laptop dataset for all the STL and MTL experiments, in order to reduce hyperparameter search. We follow the result checklist for hyperparameter searches from (Dodge et al., 2019), details of which can be found in tables 17 and 18 of Appendix F along with Figure 2 showing the expected validation scores from the hyperparameter tuning. For the MTL model, a single epoch involves training for one epoch on the auxiliary task and then an epoch on the main task, as previous work has shown training the lower-level task first improves overall results (Hashimoto et al., 2017). In this work, we assume all of the auxiliary training tasks are conceptually lower than TSA.

Evaluation: For all experiments, we run each model five times (Reimers and Gurevych, 2017) and report the mean and standard derivation. We also take the distribution of the five runs to perform significance testing (Reimers and Gurevych, 2018). Following Dror et al. (2018), we use the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945) for the $F_1$ metrics and a more powerful parametric Welch’s $t$-test (Welch, 1947) for the accuracy metric.

4.1 Results

We report the $F_1$ score for the target extraction ($F_{1-a}$), macro $F_1$ ($F_{1-s}$) and accuracy score ($acc-s$) for the sentiment classification for all targets that have been correctly identified by the model, and finally the $F_1$ score for the full targeted task ($F_{1-i}$), following He et al. (2019). We do not report the $F_1$ for opinion term extraction ($F_{1-o}$) as we do not perform opinion term extraction. Unlike He et al. (2019), we do not use any of the samples that contain the conflict label in the Laptop or Restaurant datasets. The test split results for the main $F_{1-i}$ metric are reported in Table 4 and for conciseness, the other metrics for the test split are reported in tables 9 and 10 of Appendix C.

As shown in Table 4, on the $F_{1-i}$ metric the STL model performs best on four of the eight experiments, although it is only significantly better than the majority of MTL models for the CWR model on MPQA. Of the MTL models, MTL (CD) + GloVe performs best on MPQA (18.88), MTL (DR) + GloVe is best on Restaurant (66.06), and MTL (LEX) is the best model on Laptop (54.85) with GloVe and Restaurant (71.77) with CWR. The CWR models consistently outperform the GloVe models – by an average of 5.4

\^1\text{For this Transformer ELMo it uses the output from the 6 transformer layers and the output from the non-contextualised character encoder, thus in total 7 layers are weighted and summed.}
percentage points (pp) across all experiments – and give the best performance on all datasets.

The results suggest that MTL has a greater impact when using the GloVe embeddings, rather than transfer learning (CWR). It is, however, unclear when it is beneficial to combine transfer learning and MTL, as only on the Restaurant datasets an MTL model improves over the STL when using transfer learning, as no MTL + CWR performs better than its STL counterpart. However the results also show that MTL does not hurt the STL models, as no STL model is significantly better than all of the MTL models across the datasets and embeddings for the $F_{1i}$ metric. These findings also generalise to the results on the development splits, which can be found in tables 11 and 12 within Appendix C.

$$\begin{array}{|l|c|c|c|c|c|}
\hline
\text{Laptop} & \text{MAMS} & \text{Restaurant} & \text{MPQA} \\
\hline
\text{GloVe} & \text{CWR} & \text{GloVe} & \text{CWR} & \text{GloVe} & \text{CWR} \\
\hline
\text{MTL (CD)} & 54.65 & 62.89 & 62.50 & 65.17 & 65.06 & 71.04 \\
& (1.37) & (1.18) & (0.42) & (0.35) & (2.66) & (1.13) \\
\text{MTL (DR)} & 53.67 & 62.29 & 62.05* & 65.10 & 66.06 & 71.45 \\
& (0.94) & (1.32) & (0.32) & (0.63) & (2.63) & (1.47) \\
\text{MTL (LEX)} & \textbf{54.85} & 62.55 & 62.14* & \textbf{64.65}* & 65.89 & \textbf{71.77} \\
& (0.99) & (1.66) & (0.83) & (0.88) & (1.32) & (1.88) \\
\text{MTL (SFU)} & 53.73 & 62.61 & 62.34* & 65.00* & 65.82 & 71.63 \\
& (1.93) & (1.79) & (0.54) & (0.48) & (1.31) & (1.64) \\
\text{MTL (SPEC)} & 51.65 & 62.03 & 62.16 & 64.50* & 65.16 & 71.51 \\
& (2.32) & (1.14) & (0.71) & (0.79) & (1.50) & (1.16) \\
\text{MTL (UPOS)} & 54.17 & 62.35 & 62.79 & 64.88 & 65.73 & 70.38 \\
& (2.26) & (0.77) & (0.37) & (0.46) & (1.46) & (1.63) \\
\text{STL} & 54.37 & \textbf{63.70} & \textbf{63.20} & \textbf{65.70} & 65.60 & 70.68 \\
& (2.56) & (1.14) & (0.65) & (0.55) & (1.06) & (1.53) \\
\hline
\end{array}$$

Table 4: The $F_{1i}$ results for the test split, where the values represent the mean (standard deviation) of five runs with a different random seed. The bold values represent the best performing model for that dataset and embedding. The * represent the models that perform statistically significantly worse than the STL model for that dataset and embedding at a 95% confidence level.

5 Analysis

Given that the main results seem inconclusive as to what effect MTL has on TSA, we wish to perform a more detailed analysis of how the STL and MTL models differ. We first test all models trained on the original Laptop and Restaurant datasets on the LaptopNeg, RestaurantNeg, LaptopSpec, and RestaurantSpec test splits. The results for negation and speculation can be seen in tables 5 and 6 respectively. The results for the dev split and the $F_{1i}$-s of the test split can be found in Appendix D.

On LaptopNeg and RestaurantNeg, the MTL models trained with additional negation data (CD and SFU) perform an average of 3.8 pp better than the STL model on the $F_{1i}$ metric using GloVe embeddings. Furthermore, the MTL (SFU) model with GloVe embeddings on the $F_{1i}$-i and acc-s metrics significantly outperforms the STL model across both datasets and splits. While MTL of negation helps the sentiment classification scores ($acc-s$) on LaptopNeg and RestaurantNeg, it does not help extraction.

This makes sense conceptually, as negation has little effect on whether or not a word is part of a sentiment target. Instead, jointly learning dependency relations (DR) and full lexical analysis (LEX) give the best extraction results. Further we find that on average the best MTL model (MTL (SFU)) when combined with transfer learning does marginally beat the STL equivalent.

While the speculation results do find MTL models improves results when using GloVe embeddings, with the additional speculation (SPEC) and dependency relation (DR) data improving the $F_{1i}$ metric by 0.5 pp and 0.49 pp respectively on average. However only on the Restaurant dataset does the MTL still improve results when using CWR. Unlike the negation data results, the speculation results appear to be helped more by syntactic auxiliary tasks like DR than semantic tasks like CD and to some extent SFU.

See the additional python notebooks that accompany the code we release to show empirically, that for the LaptopNeg development results MTL (SFU) is significantly better than the STL model using the GloVe embedding.
highlighted models are those of MTL. We show that TSA methods are drastically affected by negation and speculation for TSA and the usefulness of the created a negation and speculation annotated challenge dataset.

In this paper, we have compared the effects of MTL using various auxiliary tasks for TSA and have found that all models perform comparatively worse on the challenge datasets, dropping an average of 24 and 25 pp on \( F_1 - i \) on the negation and speculation data, respectively. Nearly all of this drop comes from poorer classification (\( acc-s, F_1 - s \)), while target extraction (\( F_1 - a \)) is relatively stable. This demonstrates the importance of resolving negation and speculation for TSA and the usefulness of the annotated data to determine these effects.

### 6 Conclusion

In this paper, we have compared the effects of MTL using various auxiliary tasks for TSA and have created a negation and speculation annotated challenge dataset\(^\text{13}\) for TSA in order to isolate the effects of MTL. We show that TSA methods are drastically affected by negation and speculation effects in the

\(^\text{13}\)The development \( F_1 - i \) result for MTL (LEX) on the Laptop dataset is worse than STL by 0.05 but for all other \( F_1 - i \) Laptop results MTL (LEX) is better than STL.

\(^\text{14}\)Dataset can be found at [https://github.com/jbarnesspain/multitask_negation_for_targeted_sentiment](https://github.com/jbarnesspain/multitask_negation_for_targeted_sentiment)

|                     | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|---------------------|----------|----------|-----------|-----------|------------|------------|-----|
| **sentiment**       |          |          |           |           |            |            |     |
| GloVe               | 42.80    | 38.54\*  | 38.72\*   | 45.26     | 41.23\*    | 38.92\*    | 38.32\* |
| CWR                 | 48.49    | 45.90    | 45.93     | 47.04     | 45.71      | 46.29      | 46.50 |
| **Laptop**          |          |          |           |           |            |            |     |
| GloVe               | 75.36\*  | 76.05\*  | 78.68     | 75.04\*   | 76.14      | 77.98      | 76.52\* |
| CWR                 | 82.39    | 82.95    | 83.47     | 82.24\*   | 82.58      | 82.10      |      |
| **targeted**        |          |          |           |           |            |            |     |
| GloVe               | 32.28    | 29.30\*  | 30.47\*   | 33.96     | 31.36\*    | 30.36\*    | 29.33\* |
| CWR                 | 39.95    | 38.08    | 38.35     | 39.18     | 37.59      | 38.23      | 38.14 |
| **Restaurant**      |          |          |           |           |            |            |     |
| GloVe               | 53.41    | 49.78\*  | 47.69\*   | 56.01     | 48.86\*    | 50.58\*    | 49.86\* |
| CWR                 | 60.69    | 62.61    | 60.80     | 60.45     | 61.70      | 60.06      | 60.66 |
| **extraction**      |          |          |           |           |            |            |     |
| GloVe               | 80.97    | 82.22    | 82.15     | 80.74     | 81.53      | 81.92      | 80.97 |
| CWR                 | 83.04    | 82.94\*  | 84.10     | 83.94     | 83.48      | 82.33\*    | 83.50 |
| **targeted**        |          |          |           |           |            |            |     |
| GloVe               | 43.28    | 40.95\*  | 39.19\*   | 45.22     | 39.85\*    | 41.43\*    | 40.38\* |
| CWR                 | 50.40    | 51.92    | 51.15     | 50.75     | 51.49      | 49.45      | 50.68 |

Table 5: Sentiment (\( acc-s \)), extraction (\( F_1 - a \)) and full targeted (\( F_1 - i \)) results for Laptop\(_{neg}\) and Restaurant\(_{neg}\) test split, where the values represent the mean (standard deviation) of five runs with a different random seeds. The **bold** values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represents the models that are significantly worse (\( p < 0.05 \)) than the best performing model on the respective dataset, metric, and embedding.
highlighted models are those beneficial auxiliary tasks, whereby this approach could lead to MTL and transfer learning better com-

Future work should consider using a pipeline approach, where each subtask is paired with the most beneficial auxiliary tasks, whereby this approach could lead to MTL and transfer learning better com-

Table 6: Sentiment \((\text{acc-s})\), extraction \((F_1-a)\) and full targeted \((F_1-i)\) results for Laptop\textsuperscript{Spec} and Restaurant\textsuperscript{Spec} test split, where the values represent the mean \((\text{standard deviation})\) of five runs with a different random seeds. The **bold** values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represents the models that are significantly worse \((p < 0.05)\) than the best performing model on the respective dataset, metric, and embedding.

|                | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|----------------|-----------|-----------|-----------|-----------|------------|------------|------|
| **sentiment**  |           |           |           |           |            |            |      |
| GloVe          | 34.32     | **35.67** | **36.75** | **35.98** | **36.74**  | 35.57      | 34.67 |
| CWR            | 35.42     | 34.76     | 35.06     | 34.08*    | 35.03      | 35.01      | **35.97** |
| **extraction** |           |           |           |           |            |            |      |
| GloVe          | 74.77*    | 74.01*    | **77.80** | 75.99     | 73.39*     | 76.80      | 75.01 |
| CWR            | **80.11** | **80.77** | **81.47** | **83.14** | **81.49**  | **81.07**  | **79.84** |
| **targeted**   |           |           |           |           |            |            |      |
| GloVe          | 25.67*    | **26.39** | **28.59** | 27.33     | 26.95      | 27.31      | 26.01* |
| CWR            | 28.36     | 28.09     | 28.56     | 28.33     | 28.54      | 28.37      | **28.72** |

Table 6: Sentiment \((\text{acc-s})\), extraction \((F_1-a)\) and full targeted \((F_1-i)\) results for Laptop\textsuperscript{Spec} and Restaurant\textsuperscript{Spec} test split, where the values represent the mean \((\text{standard deviation})\) of five runs with a different random seeds. The **bold** values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represents the models that are significantly worse \((p < 0.05)\) than the best performing model on the respective dataset, metric, and embedding.

As the results from the standard datasets found using MTL does not always improve performance and results from \(F_1-i\) on Laptop\textsuperscript{Spec} Table \([6]\) showed that transfer learning can harm performance.\footnote{Compare the performance of MTL (LEX) using GloVe (28.59) to when it uses CWR (28.56).} Additionally the results from the challenge datasets showed that different auxiliary tasks improved the performance of different subtasks of TSA through the extraction and sentiment classification metrics. This may suggest that the target extraction and sentiment classification tasks should not be treated as a collapsed labelling task, as the sentiment and extraction tasks are not similar enough \((\text{Hu et al., 2019})\). Future work should consider using a pipeline approach, where each subtask is paired with the most beneficial auxiliary tasks, whereby this approach could lead to MTL and transfer learning better com-
plimenting each other. Finally, we release the code[^16] dataset, and trained models associated with this paper, hyperparameter search details with compute infrastructure (Appendix E), number of parameters and runtime details (Appendix F), and further detailed dev and test results (appendices C and D), in line with the result checklist from [Dodge et al. (2019)].

**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). Andrew has been funded by Lancaster University by an EPSRC Doctoral Training Grant. The authors thank the UCREL research centre for hosting the models created from this research.

**References**

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, BC, Canada, October.

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*, pages 1896–1906, New Orleans, USA.

Jimmy Ba and Rich Caruana. 2013. Do deep nets really need to be deep? In *NIPS*.

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, August.

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2020. Improving sentiment analysis with multitask learning of negation. *Natural Language Engineering*, 27.

Jeremy Barnes. 2019. Ltg-oslo hierarchical multi-task network: The importance of negation for document-level sentiment in spanish. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 378–389.

Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–725. ACM.

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.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Philipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the International AAAI Conference on the Web and Social Media*.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! Born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy, July. Association for Computational Linguistics.

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.

Isaac Councill, Ryan McDonald, and Leonid Velikovich. 2010. What’s great and what’s not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden, July.

[^16]: [https://github.com/jbarnesspain/multitask_negation_for_targeted_sentiment]
Noa P. Cruz, Maite Taboada, and Ruslan Mitkov. 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

Sanjiv R. Das and Mike Y. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, September.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Shiao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khaskabi, Kevin Lin, Jiannming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Alky Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 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. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheong Li, and Yiwei Lv. 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. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruiying Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6279–6284, Hong Kong, China, November. Association for Computational Linguistics.

Wei Jin and Hung Hay Ho. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, page 465–472, New York, NY, USA. Association for Computing Machinery.

Arzoo Katiyar and Claire Cardie. 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. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. 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 647–653, Minneapolis, Minnesota, June. Association for Computational Linguistics.
Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia event extraction shared task, 2013 edition - overview. In Proceedings of the BioNLP Shared Task 2013 Workshop, pages 8–15, Sofia, Bulgaria, August. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2013. Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 848–854.

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 8th International Conference on Language Resources and Evaluation, pages 3190–3195, Istanbul, Turkey.

Meghana Kshirsagar, Sam Thomson, Nathan Schneider, Jaime Carbonell, Noah A. Smith, and Chris Dyer. 2015. Frame-semantic role labeling with heterogeneous annotations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 218–224, Beijing, China, July.

Emanuele Lapponi, Jonathon Read, and Lilja Øvrelid. 2012. Representing and resolving negation for sentiment analysis. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, pages 687–692, Washington, DC, USA.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 34–41, Hong Kong, China, November. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, pages 2873–2879, New York, USA.

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

Dehong Ma, Sujuan Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4737–4742, Brussels, Belgium, October-November. Association for Computational Linguistics.

Ana Marasović and Anette Frank. 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.

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, pages 44–53, Valencia, Spain.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’15, page 43–52, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations, pages 1–12.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1643–1654, Seattle, Washington, USA, October. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey.

Lilja Øvrelid, Petter Mahlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for norwegian. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 5025–5033, Marseille, France, May. European Language Resources Association.
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, August.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 1532–1543, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. 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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1499–1509, Brussels, Belgium, October-November. Association for Computational Linguistics.

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

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud Maria Jiménez-Zafra, and Gülşen Eryiğit. 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.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv preprint arXiv:1803.09578.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In Association for Computational Linguistics (ACL).

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, USA.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. Language Resources and Evaluation, 43:227–268.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1537–1547, Denver, Colorado, May–June. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jy Wu, Jason Chuang, Chris Manning, Andrew Ng, and Chris 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, 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.
Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, Suppl 11.

Wenya Wang and Sinno Jialin Pan. 2018a. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2171–2181, Melbourne, Australia, July. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2018b. Transition-based adversarial network for cross-lingual aspect extraction. In *IJCAI*, pages 4475–4481.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017. TDParse: Multi-target-specific sentiment recognition on twitter. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 483–493, Valencia, Spain, April. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

B. L. Welch. 1947. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015a. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 612–621, September.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015b. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal, September. Association for Computational Linguistics.

Meishan Zhang, Peili Liang, and Guohong Fu. 2019. Enhancing opinion role labeling with semantic-aware word representations from semantic role labeling. 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 641–646, Minneapolis, Minnesota, June. Association for Computational Linguistics.
Appendix

Appendix A  Class Distribution of the Sentiment Datasets

|          | Train   | Dev     | Test     |
|----------|---------|---------|----------|
|          | pos     | neu     | neg      | both    | pos     | neu     | neg      | both    | pos     | neu     | neg      | both    |
| Laptop   | 19.9    | 43.2    | 36.9     | -       | 18.0    | 40.6    | 41.4     | -       | 26.0    | 53.5    | 20.5     | -       |
| Laptop\_Neg | -       | -       | -       | -       | -       | 17.1    | 35.4     | 47.5     | -       | 26.7    | 23.3    | 50.0     | -       |
| Laptop\_Spec | -       | -       | -       | -       | 50.7    | 16.2    | 33.1     | -       | 38.2    | 20.5    | 41.4     | -       |
| Restaurant | 15.8    | 60.0    | 24.2     | -       | 12.3    | 65.2    | 22.5     | -       | 11.5    | 66.6    | 21.9     | -       |
| Restaurant\_Neg | -       | -       | -       | -       | 16.4    | 32.5    | 51.1     | -       | 15.0    | 32.2    | 52.8     | -       |
| Restaurant\_Spec | -       | -       | -       | -       | 30.0    | 29.0    | 41.0     | -       | 16.9    | 39.7    | 43.5     | -       |
| MAMS     | 45.1    | 30.2    | 24.7     | -       | 45.5    | 30.3    | 24.3     | -       | 45.5    | 29.9    | 24.6     | -       |
| MPQA     | 13.3    | 43.9    | 39.1     | 3.7     | 17.0    | 42.5    | 37.0     | 3.5     | 19.2    | 33.2    | 41.4     | 6.3     |

Table 7: Sentiment class distribution statistics as a percentage of the number of targets (samples), for the sentiment datasets used in the experiments. pos, neu, neg, and both represent the sentiment classes positive, neutral, negative, and both respectively.

Appendix B  Examples of Auxiliary Tasks

|          | you    | might   | not     | like    | the     | service |
|----------|--------|---------|---------|---------|---------|---------|
| CD       | B\_scope | I\_scope | B\_cue | B\_scope | I\_scope | I\_scope |
| SFU      | B\_scope | I\_scope | B\_cue | B\_scope | I\_scope | I\_scope |
| SPEC     | B\_scope | B\_cue | B\_scope | I\_scope | I\_scope | I\_scope |
| UPOS     | PRON   | AUX     | PART    | VERB    | DET     | NOUN    |
| DR       | nsubj  | aux     | advmod  | root    | det     | obj     |
| LEX      | O\_PRON | O\_AUX | O\_ADV | B\_v\_emotion | O\_DET | B\_N\_n\_ACT |

Table 8: A toy example sentence with the labels from each auxiliary task
### Appendix C Additional Main Result Tables

|                  | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL          |
|------------------|----------|----------|-----------|-----------|------------|------------|--------------|
| acc-s            |          |          |           |           |            |            |              |
| GloVe            | **71.90**| 70.66    | 70.36     | 70.30     | 68.11*     | 69.60*     | 70.80        |
|                  | (1.32)   | (1.55)   | (2.24)    | (1.86)    | (2.19)     | (1.98)     | (2.02)       |
| CWR              | 75.30    | 74.75    | 74.56     | 74.36     | 74.47*     | 74.70*     | **76.85**    |
|                  | (0.54)   | (1.14)   | (1.49)    | (1.74)    | (0.82)     | (1.02)     | (1.96)       |
| F1-s             |          |          |           |           |            |            |              |
| GloVe            | **65.00**| 63.19    | 63.07     | 62.60*    | 59.83*     | 66.63*     | **69.91**    |
|                  | (1.36)   | (2.32)   | (3.51)    | (2.74)    | (2.46)     | (1.61)     | (3.32)       |
| CWR              | 66.92    | 67.76    | 67.26     | 66.00     | 66.92*     | 66.63*     | **69.91**    |
|                  | (2.41)   | (1.75)   | (2.27)    | (3.03)    | (1.21)     | (1.61)     | (2.72)       |
| extraction       |          |          |           |           |            |            |              |
| GloVe            | 76.00*   | 75.98    | **77.99** | 76.43     | 75.81      | 77.81      | 76.76        |
|                  | (0.99)   | (1.17)   | (1.14)    | (1.57)    | (1.57)     | (1.57)     | (1.09)       |
| CWR              | 83.51    | 83.32    | 83.88     | **84.21** | 83.29      | 83.48      | 82.90        |
|                  | (1.09)   | (0.94)   | (0.88)    | (1.81)    | (1.15)     | (1.30)     | (0.72)       |
| targeted         |          |          |           |           |            |            |              |
| GloVe            | 54.65    | 53.67    | **54.85** | 53.73     | 51.65*     | 54.17      | 54.37        |
|                  | (1.37)   | (0.94)   | (0.99)    | (1.93)    | (2.32)     | (2.26)     | (2.56)       |
| CWR              | 62.89    | 62.29    | 62.55     | 62.61     | 62.03      | 62.35      | **63.70**    |
|                  | (1.18)   | (1.32)   | (1.66)    | (1.79)    | (1.14)     | (0.77)     | (1.14)       |
| acc-s            |          |          |           |           |            |            |              |
| GloVe            | **78.18**| 74.37    | 75.94     | 77.38     | 72.82*     | **73.83**  | 73.01*       |
|                  | (2.72)   | (3.47)   | (4.91)    | (3.88)    | (3.88)     | (3.41)     | (3.41)       |
| CWR              | 71.84    | 72.01    | **73.08** | 70.96     | 72.79      | 72.61      | 70.47        |
|                  | (3.46)   | (3.01)   | (4.01)    | (2.03)    | (3.13)     | (3.84)     | (1.51)       |
| F1-s             |          |          |           |           |            |            |              |
| GloVe            | **42.03**| 39.96    | 40.58     | 41.16     | 39.19      | **39.90**  | 39.00        |
|                  | (1.50)   | (1.66)   | (2.32)    | (1.94)    | (1.06)     | (1.06)     | (1.86)       |
| CWR              | 39.92    | 40.27    | **41.13** | 39.17     | 39.84      | **39.90**  | 39.25        |
|                  | (1.15)   | (0.80)   | (3.05)    | (0.79)    | (1.56)     | (1.66)     | (0.68)       |
| extraction       |          |          |           |           |            |            |              |
| GloVe            | 24.17    | 22.93*   | 24.58     | 22.84*    | 22.90*     | **25.34**  | 24.77        |
|                  | (1.44)   | (1.57)   | (1.36)    | (1.58)    | (2.49)     | (3.55)     | (3.55)       |
| CWR              | 30.98*   | 30.71*   | 31.19*    | 31.41*    | 31.41*     | **34.99**  | **34.99**    |
|                  | (2.40)   | (1.10)   | (2.69)    | (1.16)    | (0.51)     | (2.62)     | (1.10)       |
| targeted         |          |          |           |           |            |            |              |
| GloVe            | **18.88**| 17.03*   | **18.66** | 17.60*    | 16.70*     | **18.70**  | 18.11        |
|                  | (1.17)   | (1.12)   | (1.22)    | (0.57)    | (2.26)     | (0.25)     | (2.83)       |
| CWR              | 22.25*   | 22.09*   | 22.74     | 22.30*    | 22.86*     | **23.05**  | **24.66**    |
|                  | (2.06)   | (0.70)   | (1.68)    | (1.99)    | (0.98)     | (0.88)     | (1.07)       |

Table 9: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for Laptop and MPQA test split, where the values represent the mean (standard deviation) of five runs with a different random seed. The **bold** values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and CWR at a 95% confidence level.
|          | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|----------|----------|----------|----------|----------|------------|------------|-----|
| acc-s    | 81.72 (1.12) | 81.42 (1.02) | 81.24 (0.44) | 81.10 (0.46) | 80.90 (0.76) | 81.58 (0.96) | 81.70 (0.79) |
| MAMS     | 84.59 (0.50) | 84.81 (0.84) | 83.99* (0.84) | 83.73 (0.16) | 84.28 (0.94) | 83.90* (0.51) | 84.67 (0.56) |
| F1-s     | 81.05 (1.15) | 80.71 (0.43) | 80.53 (0.69) | 80.39 (0.80) | 80.81 (0.67) | 80.94 (0.88) | 80.94 (0.88) |
| extraction | 76.49 (0.99) | 76.21* (0.39) | 76.49 (1.06) | 76.87 (0.64) | 76.97 (0.48) | 77.36 (0.54) | 77.19 (0.19) |
| CWR      | 77.04 (0.35) | 76.76* (0.13) | 76.98 (0.84) | 77.64 (0.36) | 76.54* (0.79) | 77.33 (0.61) | 77.59 (0.35) |
| targeted | 62.50 (0.42) | 62.05* (0.32) | 62.14* (0.83) | 62.34* (0.54) | 62.16 (0.71) | 62.79 (0.37) | 63.20 (0.65) |
| CWR      | 65.17 (0.35) | 65.10 (0.63) | 64.65* (0.88) | 65.00* (0.48) | 64.89 (0.79) | 65.70 (0.55) | 65.70 (0.55) |

|          | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|----------|----------|----------|----------|----------|------------|------------|-----|
| acc-s    | 83.02 (1.32) | 83.23 (1.09) | 83.26 (0.99) | 83.80 (0.78) | 83.01 (1.16) | 83.36 (1.09) | 83.65 (0.48) |
| Restaurant | 87.40 (0.67) | 87.63 (0.76) | 87.37 (0.90) | 87.26 (0.96) | 87.36 (0.48) | 87.00* (0.56) | 87.32 (0.66) |
| F1-s     | 66.75 (3.75) | 67.79 (3.00) | 67.59 (1.39) | 67.75 (1.92) | 67.35 (3.02) | 67.13 (2.31) | 68.00 (1.61) |
| CWR      | 72.27 (1.14) | 72.96 (1.79) | 73.73 (2.60) | 72.12 (2.30) | 73.90 (2.82) | 71.61 (1.13) | 73.47 (1.10) |
| extraction | 78.33 (1.55) | 79.34 (1.60) | 79.13 (0.93) | 78.53 (0.96) | 78.48 (0.81) | 78.84 (0.78) | 78.42 (0.85) |
| CWR      | 81.27 (0.90) | 81.53 (1.01) | 82.13 (1.35) | 82.08 (1.09) | 81.85 (1.22) | 80.89* (1.37) | 80.94 (1.18) |
| targeted | 65.06 (2.66) | 66.06 (2.63) | 65.89 (1.32) | 65.82 (1.31) | 65.16 (1.15) | 65.73 (1.46) | 65.60 (1.06) |
| CWR      | 71.04 (1.13) | 71.45 (1.47) | 71.77 (1.88) | 71.63 (1.64) | 71.51 (1.16) | 70.38 (1.63) | 70.68 (1.53) |

Table 10: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for MAMS and Restaurant test split, where the values represent the mean (standard deviation) of five runs with a different random seed. The **bold** values represent the best model, while **highlighted** models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and CWR at a 95% confidence level.
|             | acc-s | F1-s | extraction | targeted |
|-------------|-------|------|------------|----------|
| Laptop      |       |      |            |          |
| MTL (CD)    | 77.47 | 71.55| 57.99      | 65.62*   |
| MTL (DR)    | 77.53 | 70.75| 70.57      | 64.42    |
| MTL (LEX)   | 76.52 | 70.61| 75.70      | 64.73    |
| MTL (SFU)   | 78.34 | 70.94| 58.62      | 65.52*   |
| MTL (SPEC)  | 77.32 | 70.70| 57.80      | 64.72*   |
| MTL (UPOS)  | 76.67 | 67.82| 56.82      | 64.55*   |
| STL         | 77.22 | 67.27| 57.97      | 66.51    |
| CWR         | 81.22 | 79.63*| 80.82      | 75.99*   |
|             | 77.53 | 74.84 | 80.89      | 57.02    |
|             | 76.52 | 74.89 | 80.39      | 57.92    |
|             | 78.34 | 73.70 | 81.86      | 57.92    |
|             | 77.32 | 75.22 | 80.05*     | 58.62    |
|             | 76.67 | 71.90 | 81.37      | 56.82    |
|             |       |       |            | 57.97    |
| MPQA        |       |      |            |          |
| MTL (CD)    | 87.75 | 88.65| 54.18      | 20.68    |
| MTL (DR)    | 88.63 | 90.08| 58.50      | 18.57    |
| MTL (LEX)   | 87.64 | 87.23| 59.08      | 18.35    |
| MTL (SFU)   | 89.11 | 85.62| 59.03      | 18.23    |
| MTL (SPEC)  | 86.85 | 85.62| 56.55      | 18.23    |
| MTL (UPOS)  | 88.16 | 88.71| 53.82      | 18.23    |
| STL         | 85.29*| 88.75| 55.74      | 18.23    |
| CWR         | 88.63 | 90.08| 52.83*     | 20.68    |
|             | 87.64 | 87.23| 58.50      | 18.35    |
|             | 89.11 | 85.62| 59.03      | 18.23    |
|             | 86.85 | 85.62| 56.55      | 18.23    |
|             | 88.16 | 88.71| 53.82      | 18.23    |
|             |       |       |            | 18.23    |

Table 11: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for Laptop and MPQA development split, where the values represent the mean (standard deviation) of five runs with a different random seed. The **bold** values represent the best model, while **highlighted** models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and CWR at a 95% confidence level.
|          | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL       |
|----------|----------|----------|-----------|-----------|------------|------------|-----------|
| acc-s    | GloVe    | 80.73*   | 80.98     | 80.89*    | 80.68*     | 80.87*     | 81.36     |
|          |          | (0.22)   | (1.17)    | (0.45)    | (0.48)     | (0.53)     | (0.82)    |
|          | CWR      | 84.37    | 83.99*    | **84.73** | 84.18*     | 84.17*     | 83.79*    |
|          |          | (0.24)   | (0.25)    | (0.34)    | (0.49)     | (0.49)     | (0.40)    |
| F1-s     | GloVe    | 80.20*   | 80.48*    | 80.38*    | 80.14*     | 80.38*     | 80.76     |
|          |          | (0.24)   | (1.15)    | (0.30)    | (0.57)     | (0.56)     | (0.95)    |
|          | CWR      | **84.14**| **83.72** | **84.46** | 83.96*     | 83.93*     | **85.4**  |
|          |          | (0.24)   | (0.21)    | (0.35)    | (0.47)     | (0.47)     | (0.42)    |
| extraction | GloVe   | 78.93    | 79.11     | **79.24** | 79.00      | 78.86      | 78.68     |
|          |          | (0.66)   | (0.52)    | (0.62)    | (0.47)     | (0.67)     | (0.35)    |
|          | CWR      | 77.81    | 77.86     | 77.62*    | 78.32      | 77.59*     | **78.54** |
|          |          | (0.48)   | (0.59)    | (0.34)    | (0.31)     | (0.21)     | (0.38)    |
| targeted | GloVe    | 63.72*   | 64.06*    | 64.10*    | 63.74*     | 63.76*     | 64.01*    |
|          |          | (0.63)   | (0.66)    | (0.48)    | (0.28)     | (0.21)     | (0.62)    |
|          | CWR      | 65.65    | 65.39*    | 65.77*    | 65.93      | 65.31*     | 65.81     |
|          |          | (0.39)   | (0.47)    | (0.44)    | (0.20)     | (0.56)     | (0.41)    |
| Restaurant | GloVe   | 78.42*   | 78.78     | **79.58** | 78.75      | 78.31*     | **79.14** |
|          |          | (0.78)   | (0.67)    | (0.89)    | (0.52)     | (0.78)     | (0.78)    |
|          | CWR      | **81.90**| **81.90** | 81.53     | **81.89**  | 81.02      | 80.47*    |
|          |          | (0.69)   | (0.69)    | (0.86)    | (0.84)     | (0.85)     | (1.10)    |
| F1-s     | GloVe    | 62.89    | 64.01     | **65.37** | 62.49*     | 62.54*     | 63.00*    |
|          |          | (2.84)   | (2.56)    | (1.47)    | (0.86)     | (2.28)     | (1.64)    |
|          | CWR      | 67.98    | 69.26     | 69.09     | 68.18      | 67.54      | 67.14*    |
|          |          | (3.46)   | (1.07)    | (1.72)    | (1.90)     | (3.88)     | (1.05)    |
| extraction | GloVe   | 78.22*   | **79.20** | 78.85     | 78.38*     | 78.21      | **79.62** |
|          |          | (0.78)   | (1.07)    | (1.08)    | (0.73)     | (1.37)     | (0.48)    |
|          | CWR      | 81.69    | 81.84     | **82.56** | 82.25      | 82.07      | **82.48** |
|          |          | (0.71)   | (0.88)    | (1.79)    | (0.22)     | (0.68)     | (0.61)    |
| targeted | GloVe    | 61.34*   | 62.39     | 62.75     | 61.73*     | 61.25*     | **63.02** |
|          |          | (0.73)   | (0.95)    | (1.26)    | (0.77)     | (1.26)     | (0.42)    |
|          | CWR      | 66.90    | 67.03     | 67.31     | **67.36**  | 66.49      | 66.38*    |
|          |          | (0.40)   | (1.06)    | (0.32)    | (0.66)     | (0.52)     | (1.31)    |

Table 12: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for MAMS and Restaurant development split, where the values represent the mean (standard deviation) of five runs with a different random seed. The bold values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and embedding at a 95% confidence level.
### Appendix D  Additional Negation and Speculation Result Tables

|                | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|----------------|----------|----------|-----------|-----------|------------|------------|------|
| **Laptop\textsubscript{Neg}** |          |          |           |           | **42.33**  | **39.46**  |      |
| F1-s           | GloVe    | 40.88    | 38.11*    | 38.89     | 39.46*     | 38.11*     |      |
|                |          | (2.17)   | (1.71)    | (3.32)    | (2.98)     | (2.02)     |      |
|                | CWR      | 44.81    | 45.05     | 44.58     | 43.53      | 43.83      | 44.77 |
|                |          | (2.40)   | (3.47)    | (2.29)    | (2.74)     | (1.90)     | (1.54)|
| **CWR**        | 45.08    |          |           |           |            |            |      |
|                |          |          |           |           |            |            |      |
| **Restaurant\textsubscript{Neg}** |          |          |           |           |            |            |      |
| F1-s           | GloVe    | 46.58*   | 44.16*    | 42.74*    | **47.65**  | 44.00      | 44.78 |
|                |          | (3.24)   | (2.18)    | (1.09)    | (1.35)     | (3.99)     | (1.85)|
|                | CWR      | 52.85*   | 54.41     | 54.08     | 52.54*     | **55.63**  | 52.16 |
|                |          | (1.69)   | (1.51)    | (3.87)    | (1.99)     | (1.65)     | (2.01)|
| **CWR**        | 53.59*   |          |           |           |            |            |      |
|                |          |          |           |           |            |            |      |

Table 13: \(F_1\text{-s}\) results for the **negation test split**, where the values represent the mean (standard deviation) of five runs with a different random seed. The **bold** values represent the best model, while **highlighted** models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and embedding at a 95% confidence level.

|                | MTL (CD) | MTL (DR) | MTL (LEX) | MTL (SFU) | MTL (SPEC) | MTL (UPOS) | STL |
|----------------|----------|----------|-----------|-----------|------------|------------|------|
| **Laptop\textsubscript{Spec}** |          |          |           |           |            |            |      |
| F1-s           | GloVe    | 32.74    | 33.14     | **35.24** | **33.62**  | **33.83**  | **33.21** |
|                |          | (2.35)   | (0.98)    | (2.16)    | (2.59)     | (1.62)     | (1.00) |
|                | CWR      | 33.02    | 33.33     | 33.14     | 31.72*     | 33.25      | 32.71 |
|                |          | (4.07)   | (1.56)    | (2.10)    | (0.88)     | (2.14)     | (1.38) |
| **CWR**        | **34.08**|          |           |           |            |            |      |
|                |          |          |           |           |            |            |      |
| **Restaurant\textsubscript{Spec}** |          |          |           |           |            |            |      |
| F1-s           | GloVe    | 55.27    | **57.77** | 56.27*    | **55.59**  | **57.35**  | **56.55** |
|                |          | (3.82)   | (2.91)    | (2.36)    | (1.09)     | (3.75)     | (3.14) |
|                | CWR      | 58.84*   | 60.95     | **62.36** | 58.44*     | 60.52      | 59.23 |
|                |          | (1.58)   | (0.98)    | (2.86)    | (2.24)     | (3.25)     | (1.81) |
| **CWR**        | 60.74    |          |           |           |            |            |      |
|                |          |          |           |           |            |            |      |

Table 14: \(F_1\text{-s}\) results for the **speculation test split**, where the values represent the mean (standard deviation) of five runs with a different random seed. The **bold** values represent the best model, while **highlighted** models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and embedding at a 95% confidence level.
|                  | acc-s |     | F1-s    |     | extraction |     | targeted |     |
|------------------|-------|-----|---------|-----|------------|-----|----------|-----|
|                  |       |     |         |     |            |     |          |     |
|                  | MTL   | MTL | MTL     | MTL | MTL        | MTL | STL      |     |
|                  | (CD)  | (DR)| (LEX)   | (SFU)| (SPEC)     | (UPOS)|          |     |
| acc-s            |       |     |         |     |            |     |          |     |
| GloVe            | 37.07 | 32.99* | 33.98  | 36.64 | 34.83      | 32.09* | 29.45*   |     |
|                  | (3.78)| (0.96)| (2.60)  | (1.17)| (1.09)     | (3.41) | (1.46)   |     |
| CWR              | 41.84* | 39.38* | 40.85* | 45.40 | 43.00*     | 38.77* | 41.15*   |     |
|                  | (0.54)| (2.69)| (2.80)  | (1.70)| (1.57)     | (2.64) | (3.28)   |     |
| F1-s             |       |     |         |     |            |     |          |     |
| GloVe            | 34.02 | 31.48* | 33.06  | 34.61 | 33.66      | 28.92* | 24.92*   |     |
|                  | (2.05)| (2.66)| (2.21)  | (3.34)| (1.13)     | (4.11) | (1.33)   |     |
| CWR              | 38.26* | 38.05 | 39.42  | 42.32 | 41.39      | 36.73 | 38.92    |     |
|                  | (1.67)| (3.06)| (4.02)  | (3.45)| (1.97)     | (3.45) |          |     |
| extraction       |       |     |         |     |            |     |          |     |
| GloVe            | 72.49 | 74.40 | 73.91  | 71.25*| 73.83      | 73.33 | 74.60    |     |
|                  | (1.45)| (1.64)| (0.99)  | (1.20)| (0.93)     | (1.90) | (1.51)   |     |
| CWR              | 80.94 | 81.31 | 81.21  | 81.94 | 79.00*     | 81.92 | 82.75    |     |
|                  | (1.64)| (1.29)| (1.36)  | (1.46)| (0.87)     | (1.36) | (1.80)   |     |
| targeted         |       |     |         |     |            |     |          |     |
| GloVe            | 26.87 | 24.56 | 25.12  | 26.14 | 25.71      | 23.54 | 21.98*   |     |
|                  | (2.75)| (1.21)| (1.99)  | (2.55)| (0.68)     | (2.66) | (1.26)   |     |
| CWR              | 33.86* | 32.04* | 33.14* | 37.20 | 33.98*     | 31.78* | 34.01*   |     |
|                  | (0.71)| (2.54)| (1.56)  | (1.59)| (1.47)     | (2.46) | (2.29)   |     |
| acc-s            |       |     |         |     |            |     |          |     |
| GloVe            | 46.02 | 43.13* | 41.06* | 49.02 | 41.65*     | 44.02* | 42.69*   |     |
|                  | (4.88)| (3.24)| (3.36)  | (1.31)| (4.06)     | (3.09) | (2.01)   |     |
| CWR              | 53.79 | 54.40 | 52.25  | 54.42 | 53.16      | 54.31 | 52.25    |     |
|                  | (2.56)| (3.63)| (3.63)  | (3.18)| (1.92)     | (2.49) | (2.51)   |     |
| F1-s             |       |     |         |     |            |     |          |     |
| GloVe            | 40.03 | 38.56* | 37.54  | 41.05 | 37.01      | 38.03 | 38.45    |     |
|                  | (5.30)| (3.10)| (3.69)  | (2.45)| (4.16)     | (3.49) | (2.28)   |     |
| CWR              | 48.03 | 49.13 | 48.31  | 49.18 | 48.80      | 49.42 | 48.06    |     |
|                  | (3.48)| (2.62)| (3.71)  | (3.78)| (2.03)     | (2.76) | (2.12)   |     |
| extraction       |       |     |         |     |            |     |          |     |
| GloVe            | 81.74 | 82.37 | 82.36  | 80.61*| 81.34      | 82.38 | 81.32    |     |
|                  | (0.77)| (0.64)| (0.80)  | (0.72)| (1.48)     | (1.00) | (0.37)   |     |
| CWR              | 84.08 | 82.87* | 84.32  | 83.71 | 83.45      | 84.02 | 84.84    |     |
|                  | (0.72)| (0.66)| (0.68)  | (1.09)| (1.09)     | (1.12) | (0.99)   |     |
| targeted         |       |     |         |     |            |     |          |     |
| GloVe            | 37.61 | 35.54* | 33.82* | 39.51 | 33.88*     | 36.27* | 34.71*   |     |
|                  | (3.86)| (2.81)| (2.78)  | (1.03)| (3.38)     | (2.64) | (1.58)   |     |
| CWR              | 45.23 | 45.07 | 44.04  | 45.57 | 44.38      | 45.62 | 44.31    |     |
|                  | (2.19)| (2.90)| (2.77)  | (2.91)| (2.05)     | (1.88) | (1.74)   |     |

Table 15: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for the negation development split, where the values represent the mean (standard deviation) of five runs with a different random seed. The bold values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and embedding at a 95% confidence level.
|                | acc-s | F1-s | extraction | targeted |
|----------------|-------|------|------------|----------|
| **Laptop Spec** |       |      |            |          |
| MTL (CD)       | 33.56* | 32.66 | 71.10 (1.97) | 27.96 (1.60) |
| MTL (DR)       | 35.00 | 33.99 | 69.42 (2.25) | 28.11 (1.81) |
| MTL (LEX)      | 31.70* | 30.72 | 72.61 (0.95) | 28.08 (1.81) |
| MTL (SFU)      | 34.67 | 33.20 | 70.46 (1.41) | 26.51 (1.34) |
| MTL (SPEC)     | 37.24 | 33.60 | 69.33 (1.58) | 27.16 (1.39) |
| MTL (UPOS)     | 31.59* | 32.79 | 70.58 (2.46) | 27.74 (1.31) |
| STL            | 32.57* | 34.27 | 73.04 (2.48) | 28.12 (1.32) |
| **Restaurant Spec** |       |      |            |          |
| MTL (CD)       | 35.54* | 34.73 | 80.50 (0.95) | 27.96 (1.60) |
| MTL (DR)       | 40.09 | 33.14 | 80.25 (0.90) | 28.11 (1.81) |
| MTL (LEX)      | 37.98 | 33.83 | 79.91* (0.61) | 28.08 (1.81) |
| MTL (SFU)      | 37.18 | 30.92 | 80.83 (0.61) | 26.51 (1.34) |
| MTL (SPEC)     | 37.97 | 33.36 | 79.35* (1.34) | 27.16 (1.39) |
| MTL (UPOS)     | 38.32 | 32.45 | 80.95 (0.43) | 27.74 (1.31) |
| STL            | 37.23 | 32.47 | 82.00 (1.32) | 28.12 (1.32) |

Table 16: acc-s, F1-s, extraction (F1-a) and full targeted (F1-i) results for the speculation development split, where the values represent the mean (standard deviation) of five runs with a different random seed. The bold values represent the best model, while highlighted models are those that perform better than the single task baseline. The * represent the models that are statistically significantly worse than the best performing model on the respective dataset, metric and embedding at a 95% confidence level.
Appendix E  Hyperparameter Search Space

| Hyperparameter                    | Search space     | Best assignment |
|-----------------------------------|------------------|-----------------|
| embedding                         | GloVe 300D       | GloVe 300D      |
| embedding trainable               | False            | False           |
| number of epochs                  | 150              | 150             |
| patience                          | 10               | 10              |
| metric early stopping monitored   | Span F1/F1-i     | Span F1/F1-i    |
| batch size                        | 32               | 32              |
| dropout                           | uniform-float[0, 0.5] | 0.5           |
| 1st layer LSTM hidden dimension   | uniform-integer[30, 110] | 60             |
| main task LSTM hidden dimension   | 50               | 50              |
| skip connection between embedding and main task layer | True            | True            |
| learning rate optimiser           | Adam             | Adam            |
| learning rate                     | loguniform-float[1e-4, 1e-2] | 1.5e-3         |
| gradient norm                     | 5.0              | 5.0             |
| regularisation type               | L2               | L2              |
| regularisation value              | 1e-4             | 1e-4            |

Table 17: STL search space and best assignment using the Laptop dataset.
| **GPU Infrastructure** | 1 GeForce GTX 1060 6GB GPU |
|------------------------|-----------------------------|
| **CPU Infrastructure** | AMD Ryzen 5 1600 CPU |
| **Number of search trials** | 30 |
| **Search strategy** | uniform sampling |
| **Best validation span F1/F1-i** | 0.6017 |
| **Training duration** | 18473 sec |
| **Model implementation** | [https://bit.ly/2X4ZJiH](https://bit.ly/2X4ZJiH) |

| Hyperparameter | Search space | Best assignment |
|----------------|--------------|-----------------|
| embedding      | GloVe 300D   | GloVe 300D      |
| embedding trainable | False     | False           |
| number of epochs | 150        | 150             |
| patience       | 10           | 10              |
| metric early stopping monitored | Span F1/F1-i | Span F1/F1-i |
| batch size     | 32           | 32              |
| dropout uniform-float[0, 0.5] | 0.27        |
| Shared/1st layer LSTM hidden dimension uniform-integer[30, 110] | 65 |
| main task LSTM hidden dimension | 50          | 50              |
| skip connection between embedding and main task layer | True        | True            |
| learning rate optimiser | Adam       | Adam            |
| learning rate loguniform-float[1e-4, 1e-2] | 1.9e-3      |
| gradient norm | 5.0          | 5.0             |
| regularisation type | L2          | L2              |
| regularisation value | 1e-4        | 1e-4            |

Table 18: MTL search space and best assignment using the Laptop dataset. The auxiliary task was negation detection using the Conan Doyle (CD) dataset.
Figure 2: Hyperparameter budget against expected span $F_1/F_1$-i performance for the STL and MTL models. The hyperparameter search space is stated within tables 17 and 18 for the STL and MTL models respectively. The shaded areas represent the expected performance $\pm 1$ standard deviation. Note the shaded area does not go beyond the maximum observed validation score as recommended [Dodge et al. (2019)].
## Appendix F Additional Reproducibility Statistics

| Embedding | Model       | Number of Parameters Including Auxiliary Task |   |   |
|-----------|-------------|-----------------------------------------------|---|---|
|           |             | Trainable | All             | Trainable | All             |
| GloVe     | STL         | 364,042   | 1,785,967       | 364,042   | 1,785,967       |
|           | MTL (CD)    | 385,851   | 2,403,876       | 385,122   | 2,403,147       |
|           | MTL (SFU)   | 385,851   | 7,066,176       | 385,122   | 7,065,447       |
|           | MTL (SPEC)  | 385,851   | 7,066,176       | 385,122   | 7,065,447       |
|           | MTL (UPOS)  | 388,413   | 2,952,738       | 385,122   | 2,949,447       |
|           | MTL (DR)    | 397,213   | 2,961,538       | 385,122   | 2,949,447       |
|           | MTL (LEX)   | 1,191,204 | 3,755,529       | 385,122   | 2,949,447       |
| CWR       | STL         | 1,001,170 | 56,870,931      | 1,001,170 | 56,870,931      |
|           | MTL (CD)    | 1,051,939 | 56,921,700      | 1,051,210 | 56,920,971      |
|           | MTL (SFU)   | 1,051,939 | 56,921,700      | 1,051,210 | 56,920,971      |
|           | MTL (SPEC)  | 1,051,939 | 56,921,700      | 1,051,210 | 56,920,971      |
|           | MTL (UPOS)  | 1,054,501 | 56,924,262      | 1,051,210 | 56,920,971      |
|           | MTL (DR)    | 1,063,301 | 56,933,062      | 1,051,210 | 56,920,971      |
|           | MTL (LEX)   | 1,857,292 | 57,727,053      | 1,051,210 | 56,920,971      |

Table 19: Number of parameters for each model using different embeddings ordered by number of trainable parameters. The number of parameters is different for the MTL models depending on whether the parameters from the auxiliary task are included or not, whereby the auxiliary task specific layer can be seen as the pink layer in Figure 1. The number of parameters including and not including the auxiliary task is stated as the MTL models at inference time would not use the auxiliary task parameters. As it can be seen there are a lot more trainable parameters for the MTL models ignoring the auxiliary task parameters, this is due to the hyperparameter search finding a larger shared LSTM hidden dimension to be preferable for the MTL models as can be seen from tables 17 and 18. It can be seen that for the GloVe MTL models the number of all parameters changes a lot depending on the auxiliary task, this is due to the GloVe embedding containing different number of vocabulary words, as we filter words based on those in the auxiliary and main task datasets/corpora. The large difference between the number of trainable parameters between GloVe and CWR models is due to the CWR being 724 parameters larger than the 300 parameter GloVe embeddings. Lastly the number of trainable parameters is dataset agnostic, the number of all parameters is not dataset agnostic for the GloVe models due to the vocabulary size, for clarification the model parameters reported here are for those trained on the Laptop dataset.
| Embedding | Model | Device | Batch Size | Min Time (s) | Max Time (s) |
|-----------|-------|--------|------------|--------------|--------------|
|           |       | CPU    | 1          | 10.24        | 10.45        |
|           |       |        | 8          | 7.00         | 7.21         |
|           |       |        | 16         | 6.67         | 6.91         |
|           |       |        | 32         | 6.35         | 6.51         |
| GloVe     | STL   | GPU    | 1          | 9.24         | 9.26         |
|           |       |        | 8          | 6.58         | 6.67         |
|           |       |        | 16         | 6.34         | 6.36         |
|           |       |        | 32         | 6.12         | 6.26         |
|           | MTL   | CPU    | 1          | 10.06        | 10.26        |
|           |       |        | 8          | 7.05         | 7.19         |
|           |       |        | 16         | 6.90         | 6.99         |
|           |       |        | 32         | 6.41         | 6.46         |
|           |       | GPU    | 1          | 9.43         | 9.49         |
|           |       |        | 8          | 6.60         | 6.70         |
|           |       |        | 16         | 6.26         | 6.55         |
|           |       |        | 32         | 6.10         | 6.20         |
|           |       | CPU    | 1          | 64.79        | 71.26        |
|           |       |        | 8          | 43.62        | 49.70        |
|           |       |        | 16         | 47.06        | 48.41        |
|           |       |        | 32         | 56.76        | 62.77        |
|           |       | GPU    | 1          | 23.26        | 23.79        |
|           |       |        | 8          | 8.82         | 9.09         |
|           |       |        | 16         | 8.57         | 8.86         |
|           |       |        | 32         | 8.45         | 9.78         |
|           |       | CPU    | 1          | 64.01        | 67.90        |
|           |       |        | 8          | 49.05        | 50.00        |
|           |       |        | 16         | 53.47        | 56.42        |
|           |       |        | 32         | 55.33        | 55.79        |
|           |       | GPU    | 1          | 23.81        | 23.97        |
|           |       |        | 8          | 9.19         | 9.49         |
|           |       |        | 16         | 8.54         | 8.92         |
|           |       |        | 32         | 8.43         | 8.70         |

Table 20: Run/inference times for STL and MTL models that have been trained on the Laptop dataset using either GloVe or CWR embeddings. Each model was timed in seconds (s) to generate predictions for 800 sentences, that were taken from the Laptop test split, of which this process was repeated five times and here we report the minimum (min) and maximum (max) time to generate predictions for those 800 sentences. We report these timings across different model configurations based on different batch sizes at prediction time and different devices. The trained MTL model used in this experiment was the MTL (SFU) version, this was chosen as it contains the largest number of total parameters as shown in Table 19. Further all of these times were based on the model already loaded into memory and using the Python timeit library for timings. Additionally the GPU used was a GeForce GTX 1060 6GB GPU, CPU was an AMD Ryzen 5 1600 CPU, and the computer had 16GB of RAM.