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 English-language 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 via language modelling can improve performance on these challenge datasets, but the overall performances indicate that there is still much room for improvement. We release both the datasets and the source code at https://github.com/jerbarnes/multitask_negation_for_targeted_sentiment.

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).

Modelling TSA has moved from sequence labeling using conditional random fields (CRFs) (Mitchell et al., 2013) or Recurrent Neural Networks (RNN) (Zhang et al., 2015a; Katiyar and Cardie, 2016; Ma et al., 2018), to Transformer models (Hu et al., 2019). However, all these improvements have concentrated on making the best of the relatively small task-specific datasets. As annotation for fine-grained sentiment is difficult and often has low inter-annotator agreement (Wiebe et al., 2005; Øvrelid et al., 2020), this data tends to be small and of varying quality. This lack of high-quality training data prevents TSA models from learning complex, compositional linguistic phenomena. For sentence-level sentiment classification, incorporating compositional information from relatively small amounts of negation or speculation data improves both robustness and general performance (Councill et al., 2010; Cruz et al., 2016; Barnes et al., 2020). Furthermore, transfer learning via language-modelling also improves fine-grained sentiment analysis (Hu et al., 2019; Li et al., 2019b). 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?

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

We explore a **multi-task learning** (**MTL**) approach to incorporate auxiliary task information in targeted sentiment classifiers in English in order to investigate the effects of negation and speculation in detail, we also annotate two new challenge datasets which contain negated and speculative examples. We find that the performance is negatively affected by negation and speculation, but MTL and **transfer learning** (**TL**) models are more robust than **single task learning** (**STL**). TL reduces the improvements of MTL, suggesting that TL is similarly effective at learning negation and speculation. The overall performance on the challenge datasets,

*The authors contributed equally.

however, confirms that there is still room for improvement.

The contributions of the paper are the following: i) we introduce two English challenge datasets 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 datasets, we 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, targeted sentiment analysis (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, namely extracting the opinion target and the polarity towards it.

English TSA datasets include MPQA (Wiebe et al., 2005), the SemEval Laptop and Restaurant reviews (Pontiki et al., 2014, 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) augment document sentiment datasets by asking annotators to create counterfactual examples for the IMDB dataset. More 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. Previous approaches to modelling TSA have often relied on general sequence labelling models, *e. g.* CRFs (Mitchell et al., 2013), probabilistic graphical models (Klinger and Cimiano, 2013), RNNs (Zhang et al., 2015b; Ma et al., 2018), and more recently pretrained Transformer models (Li et al., 2019b).

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,a). 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 can be an effective way to improve model performance (Collobert et al., 2011; Peng and Dredze, 2017; Martínez 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*.

Transfer learning methods (Mikolov et al., 2013; Peters et al., 2018a; Devlin et al., 2019) can leverage unlabeled data, but require training large models on large amounts of data. However, it seems even these models can be sensitive to negation (Ettinger, 2020; Ribeiro et al., 2020; Kassner and Schütze, 2020)

Specific to TSA, previous research has used MTL to incorporate document-level sentiment (He et al., 2019), or to jointly learn to extract opinion expressions (Li et al., 2019b; Chen and Qian, 2020).

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)[1] shows the distribution of the sentiment classes . We take the pre-processed Laptop and Restaurant datasets from Li et al. (2019a), and use the train, dev, and test splits that they provide. We use the NLTK word tokenizer to tokenise the Laptop, Restaurant, and MPQA datasets and Spacy 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. We convert all datasets to a targeted setup by extracting only the aspect targets and their polarity. We use the unified tagging scheme[2] following recent work (Li et al., 2019a,b) and convert all data to BIOUL format[3] 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.

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 original data with improvement on these two phenomena. Therefore, we further annotate the dev and test set for the Laptop and Restaurant datasets[4], and when possible[5], insert negation and speculation cues into sentences lacking them, 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, including those that occur naturally in the data. As this process could introduce errors regarding the polarity expressed towards the targets, we doubly annotate the polarity for 50 sentences from the original dev data, the negated dev data, and the speculation dev data and calculate Cohen's Kappa scores. The statistics and inter-annotator agreement scores (IAA) are shown in Table 1[6]. The new annotations have similarly high IAA scores (0.66-0.70) to the original data

[1] All datasets contain the following three sentiment classes positive, neutral, and negative. The MPQA dataset also includes a fourth rare class, both. Table 7 of Appendix A.

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

[3] BIOUL format tags each token as either **B**: beginning token, **I**: inside token, **O**: outside token, **U**: unit (single token), or **L**: last token.

[4] For clarification this is the SemEval 2014 Laptop dataset and the 2014, 2015, and 2016 combined Restaurant dataset.

[5] While inserting negation into new sentences is quite trivial, as one can always negate full clauses, *e. g.* It's good → It's not true that it's good, adding speculation often requires rewording of the sentence. We did not include sentences that speculation made unnatural.

[6] Table 7 of Appendix A shows the distribution of the sentiment classes.

	Train				Dev				Test				
	sents.	targs.	len.	mult.	sents.	targs.	len.	mult.	sents.	targs.	len.	mult.	IAA
Laptop	2,741	2,044	1.5	136	304	256	1.5	18	800	634	1.6	38	0.67
Laptop$_{Neg}$	–	–	–	–	147	181	1.5	41	403	470	1.6	79	0.70
Laptop$_{Spec}$	–	–	–	–	110	142	1.4	10	208	220	1.5	19	0.64
Restaurant	3,490	3,896	1.4	312	387	414	1.4	34	2,158	2,288	1.4	136	0.71
Restaurant$_{Neg}$	–	–	–	–	198	274	1.4	61	818	1,013	1.4	161	0.66
Restaurant$_{Spec}$	–	–	–	–	138	200	1.3	35	400	451	1.4	49	0.66
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.). IAA scores are reported on a subset of the data.

(0.67-0.71), confirming the quality of the annotations.

3.2 Auxiliary task data

For the multi-task learning experiments, we use six auxiliary tasks: negation scope detection using the Conan Doyle (**NEG**$_{CD}$) (Morante and Daelemans, 2012), both negation detection (**NEG**$_{SFU}$) and speculation detection (**SPEC**) on the SFU$_{NegSpec}$ dataset (Konstantinova et al., 2012), and Universal Part-of-Speech tagging (**UPOS**), Dependency Relation prediction (**DR**) and prediction of full lexical analysis (**LEX**) on the Streusle dataset (Schneider and Smith, 2015). We show the train, dev, test splits, as well as the number of labels, label entropy and label kurtosis (Martínez Alonso and Plank, 2017) 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 are shown 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 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[7] 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.

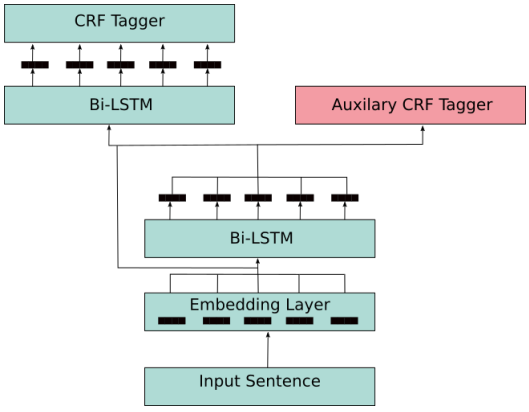

Figure 1: The overall architecture where the STL model contains all of the green layers and the MTL uses the additional pink auxiliary CRF tagger. The second Bi-LSTM has a skip connection from the embedding layer which concatenates the word embeddings with the output from the first Bi-LSTM.

Embeddings: For the embedding layer, we perform experiments using 300 dimensional **GloVe**

[7]The STL model had 1,785,967 parameters of which 364,042 were trainable as the embedding layer was frozen.

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.

	train	dev	test	# labels	label entropy	label kurtosis
NEG_{CD}	842	144	235	5	1.0	-0.8
NEG_{SFU}	13,712	1,713	1,703	5	0.2	0.2
SPEC	13,712	1,713	1,703	5	0.1	0.2
UPOS	2,723	554	535	17	2.5	-0.6
DR	2,723	554	535	49	3.1	1.3
LEX	2,723	554	535	570	3.9	75.7

Table 3: Statistics for the auxiliary datasets.

embeddings (Pennington et al., 2014), as well as **TL** from Transformer ELMo embeddings (Peters et al., 2018b)[8]. The GloVe embeddings are publicly available and trained on English Wikipedia and Gigaword data. For the MPQA dataset we use the Transformer ELMo from Peters et al. (2018b)[9] which was trained on the 1 billion word benchmark (Chelba et al., 2014). For the MAMS and Restaurant datasets we tuned a Transformer ELMo on 27 million (M) sentences from the 2019 Yelp review dataset[10], and for the Laptop dataset on 28M sentences[11] from the Amazon electronics reviews dataset (McAuley et al., 2015)[12]. Training these models on large amounts of in-domain data gives superior performance to models trained on more generic data, *e. g.* BERT (Devlin et al., 2019). For all experiments we freeze the embedding layer in order to make the results between GloVe and TL more comparable with respect to the number of trainable parameters. For TL, we learn a summed weighting of all layers[13], as this is 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 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 found in Tables 17 and 18 of Appendix E 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), eliminating the need for Bonferroni correction. 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 accu-

[8]This is a 6 layer transformer model with a bi-directional language model objective that contains 56 million parameters excluding the softmax. In comparison BERT uses a masked language modelling objective and contains 110 and 340 million parameters for the base and large versions (Devlin et al., 2019).

[9]Found at https://allennlp.org/elmo under Transformer ELMo.

[10]https://www.yelp.com/dataset

[11]More specifically there was 9M unique sentences and the model was trained for 3 epochs.

[12]For full details of on how the fine tuned Transformer ELMo models were trained see https://github.com/apmoore1/language-model.

[13]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.

		Laptop		MAMS		Restaurant		MPQA	
	Aux.	GloVe	TL	GloVe	TL	GloVe	TL	GloVe	TL
MTL	NEG_{CD}	54.65 (1.37)	62.89 (1.18)	62.50 (0.42)	65.17 (0.35)	65.06 (2.66)	71.04 (1.13)	**18.88** (1.17)	22.25* (2.00)
	DR	53.67 (0.94)	62.29 (1.32)	62.05* (0.32)	65.10 (0.63)	**66.06** (2.63)	71.45 (1.47)	17.03 (1.12)	22.09* (0.70)
	LEX	**54.85** (0.99)	62.55 (1.66)	62.14* (0.83)	64.65* (0.88)	65.89 (1.32)	**71.77** (1.88)	18.66 (1.22)	22.74 (1.68)
	NEG_{SFU}	53.73 (1.93)	62.61 (1.79)	62.34* (0.54)	65.00* (0.48)	65.82 (1.31)	71.63 (1.64)	17.60 (0.57)	22.30* (1.19)
	SPEC	51.65 (2.32)	62.03 (1.14)	62.16 (0.71)	64.50* (0.79)	65.16 (1.50)	71.51 (1.16)	16.70 (2.26)	22.86* (0.98)
	UPOS	54.17 (2.26)	62.35 (0.77)	62.79 (0.37)	64.88 (0.46)	65.73 (1.46)	70.38 (1.63)	18.70 (0.25)	23.05* (0.88)
STL		54.37 (2.56)	**63.70** (1.14)	**63.20** (0.65)	**65.70** (0.55)	65.60 (1.06)	70.68 (1.53)	18.11 (2.83)	**24.66** (1.07)

Table 4: The F_1-*i* 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.

racy 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). Unlike He et al. (2019), we do not use any of the samples that contain the *conflict* label on Laptop or Restaurant. The test results for the main F_1-*i* metric are reported in Table 4, and the other metrics for the test split are reported in Tables 9 and 10 of Appendix C.

The MTL models outperform STL on four of the eight experiments (see Table 4), although the STL TL model is significantly better than the majority of MTL models on MPQA. Of the MTL models, NEG_{CD} + GloVe performs best on MPQA (18.88), DR + GloVe is best on Restaurant (66.06), and LEX is the best model on Laptop (54.85) with GloVe and Restaurant (71.77) with TL. The TL models consistently outperform the GloVe models – by an average of 5.4 percentage points (pp) across all experiments – and give the best performance on all datasets.

The results suggest that transfer learning reduces the beneficial effects of MTL. At the same time, the results suggest 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_1-*i* metric. [14]

5 Challenge Dataset Results

In order to isolate the effects of negation and speculation on the results, we test all models trained on the original Laptop and Restaurant datasets on the Laptop_{Neg}, Restaurant_{Neg}, Laptop_{Spec}, and Restaurant_{Spec} test splits. Tables 5 and 6 show the results for negation and speculation, respectively. The results for the dev split and the F_1-*s* of the test split are shown in Appendix D.

Firstly, 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.

On Laptop_{Neg} and Restaurant_{Neg} incorporating negation auxiliary tasks gives an average improvement of 3.8 pp on the F_1-*i* metric when using GloVe embeddings. More specifically, MTL with negation improves the sentiment classification scores, but 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,

[14] These findings also generalise to the results on the development splits, shown in Tables 11 and 12 within Appendix C.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop$_{Neg}$	sentiment	GloVe	42.80 (2.48)	38.54* (0.98)	38.72* (3.00)	**45.26** (1.45)	41.23* (2.90)	38.92* (1.74)	38.32* (1.73)
		TL	**48.49** (2.32)	45.90 (3.54)	45.93 (2.13)	47.04 (2.93)	45.71 (2.19)	46.29 (2.03)	46.50 (3.30)
	extraction	GloVe	75.36* (0.91)	76.05* (1.20)	**78.68** (0.97)	75.04* (1.92)	76.14 (2.06)	77.98 (1.41)	76.52* (1.24)
		TL	82.39 (1.34)	82.95 (1.36)	**83.47** (1.26)	83.25 (1.80)	82.24* (1.39)	82.58 (1.58)	82.10 (1.11)
	targeted	GloVe	32.28 (2.23)	29.30* (0.54)	30.47* (2.45)	**33.96** (1.30)	31.36* (1.78)	30.36* (1.56)	29.33* (1.47)
		TL	**39.95** (2.02)	38.08 (3.13)	38.35 (2.01)	39.18 (2.88)	37.59 (1.99)	38.23 (1.89)	38.14 (2.23)
Restaurant$_{Neg}$	sentiment	GloVe	53.41 (4.28)	49.78* (2.10)	47.69* (1.19)	**56.01** (1.07)	48.86* (3.94)	50.58* (2.18)	49.86* (1.77)
		TL	60.69 (1.91)	**62.61** (2.11)	60.80 (3.20)	60.45 (2.04)	61.70 (1.42)	60.06 (2.13)	60.66 (2.24)
	extraction	GloVe	80.97 (1.47)	**82.22** (1.29)	82.15 (0.74)	80.74 (1.58)	81.53 (0.32)	81.92 (0.91)	80.97 (1.14)
		TL	83.04 (1.26)	82.94* (0.97)	**84.10** (0.86)	83.94 (1.67)	83.48 (1.59)	82.33* (1.37)	83.50 (1.16)
	targeted	GloVe	43.28 (3.95)	40.95* (2.31)	39.19* (1.23)	**45.22** (0.80)	39.85* (3.35)	41.43* (1.87)	40.38* (1.82)
		TL	50.40 (2.03)	**51.92** (1.64)	51.15 (3.04)	50.75 (2.10)	51.49 (0.86)	49.45 (2.01)	50.68 (2.52)

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.

jointly learning dependency relations (DR) and full lexical analysis (LEX) improve extraction results. Furthermore, when using TL instead of GloVe embeddings, the best MTL model (NEG$_{SFU}$) does marginally beat the STL TL equivalent on average, indicating that multi-task learning is still able to contribute something to transfer learning.

On Laptop$_{Spec}$ and Restaurant$_{Spec}$ MTL models improve results when using GloVe embeddings, with the additional speculation (SPEC) and dependency relation (DR) data improving the F_1-*i* metric by 0.5 pp and 0.49 pp respectively on average. However, with TL, MTL only leads to benefits on the Restaurant dataset. Unlike the negation data results, the speculation results appear to be helped more by syntactic auxiliary tasks like DR than semantic tasks like NEG$_{CD}$ and to some extent NEG$_{SFU}$.

The best MTL GloVe models on the original datasets (LEX[15] and DR, respectively) also outperform the STL GloVe models on the challenge data, indicating that MTL leads to greater robustness. When comparing the STL model using GloVe and TL on average the model improves by 9.55 pp on the negation dataset compared to 3.65 pp for the speculation suggesting that transfer learning is less effective for speculation.

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[16] 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 data. These effects can be similarly reduced by either incorporating auxiliary task information into the model through MTL or through transfer learning. Additionally, MTL of negation

[15]The development F_1-*i* result for LEX on the Laptop dataset is worse than STL by 0.05 but for all other F_1-*i* Laptop results LEX is better than STL.

[16]https://bit.ly/312kwpP

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop$_{Spec}$	sentiment	GloVe	34.32 (1.86)	35.67 (1.00)	**36.75** (1.91)	35.98 (2.05)	36.74 (1.64)	35.57 (1.31)	34.67 (1.40)
		TL	35.42 (3.54)	34.76 (1.63)	35.06 (1.97)	34.08* (0.40)	35.03 (2.36)	35.01 (1.04)	**35.97** (1.45)
	extraction	GloVe	74.77* (1.54)	74.01* (1.93)	**77.80** (1.34)	75.99 (2.48)	73.39* (1.74)	76.80 (0.99)	75.01 (1.93)
		TL	80.11* (1.40)	80.77 (1.23)	81.47* (0.50)	**83.14** (2.22)	81.49 (1.24)	81.07 (1.38)	79.84* (0.58)
	targeted	GloVe	25.67* (1.62)	26.39* (0.60)	**28.59** (1.42)	27.33 (1.70)	26.95 (1.07)	27.31 (0.82)	26.01* (1.26)
		TL	28.36 (2.81)	28.09 (1.68)	28.56 (1.60)	28.33 (0.52)	28.54 (1.83)	28.37 (0.77)	**28.72** (1.20)
Restaurant$_{Spec}$	sentiment	GloVe	62.38 (3.75)	64.01 (2.72)	63.44 (2.21)	63.33 (1.87)	**64.30** (3.14)	63.15 (3.38)	63.94 (1.84)
		TL	67.23 (1.08)	68.98 (1.17)	**69.70** (2.51)	67.62 (1.58)	66.93 (1.79)	68.13 (1.25)	68.17 (2.44)
	extraction	GloVe	75.53 (1.03)	**76.40** (1.90)	75.75 (1.18)	75.66 (1.65)	75.29 (0.77)	75.87 (0.97)	75.58 (1.48)
		TL	77.92 (1.36)	77.84 (0.84)	**79.10** (1.48)	78.76 (1.27)	78.20 (1.80)	77.15 (1.92)	77.61 (1.87)
	targeted	GloVe	47.14 (3.24)	**48.94** (3.06)	48.07 (2.22)	47.90 (1.25)	48.41 (2.48)	47.94 (3.14)	48.35 (2.32)
		TL	52.39 (1.18)	53.69 (0.69)	**55.15** (2.70)	53.25 (1.10)	52.34 (1.85)	52.55 (1.22)	52.94 (2.99)

Table 6: Sentiment (*acc-s*), extraction (F_1-*a*) and full targeted (F_1-*i*) results for Laptop$_{Spec}$ and Restaurant$_{Spec}$ 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.

can lead to small improvements when combined with transfer learning. Returning to the two original research questions, we can conclude that in general 1) MTL using negation (speculation) as an auxiliary task does make TSA models more robust to negated (speculative) samples and 2) transfer learning seems to incorporate much of the same knowledge. Additionally, incorporating syntactic information as an auxiliary task within MTL creates models that are more robust to both negation and speculation.

Neither MTL nor TL are currently guarantees for improved performance[17]. Additionally, the results from the challenge datasets indicate that different auxiliary tasks improve the performance of different subtasks of TSA. 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 too dissimilar (Hu et al., 2019). Future work should consider using pipeline or joint approaches, where each subtask can be paired with the most beneficial auxiliary tasks. This decoupling could also allow MTL and transfer learning to compliment each other more.

Finally, in order to improve reproducibility and to encourage further work, we release the code[18], 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 Lan-

[17] Compare the performance of LEX using GloVe (28.59) to when it uses TL (28.56) in Table 6 for the Laptop dataset.

[18] https://github.com/jerbarnes/multitask_negation_for_targeted_sentiment

caster 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.

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.

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.

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.

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

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, Phillipp 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 Web and Social Media*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. 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.

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.

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. 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. 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. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally 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. 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. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng 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. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng 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. 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.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

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. 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. 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. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2013. Bidirectional 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.

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. Association for Computational Linguistics.

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

Dehong Ma, Sujian 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. 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. 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 Mæhlum, 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. 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.

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. 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. 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.

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 María 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. 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. Association for Computational Linguistics.

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. 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. 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

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. 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.

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*.

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. Association for Computational Linguistics.

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. 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.

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. Association for Computational Linguistics.

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.

B Examples of Auxiliary Tasks

	you	might	not	like	the	service
NEG_{CD}	B_{scope}	I_{scope}	B_{cue}	B_{scope}	I_{scope}	I_{scope}
NEG_{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}	$\text{B}_{V-v.emotion}$	O_{DET}	$\text{B}_{N-n.ACT}$

Table 8: A toy example sentence with the labels from each auxiliary task

C Additional Main Result Tables

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop	acc-s	GloVe	**71.90** (1.32)	70.66 (1.55)	70.36 (2.24)	70.30 (1.86)	68.11* (2.19)	69.60* (1.98)	70.80 (2.02)
		TL	75.30 (0.54)	74.75 (1.14)	74.56 (1.49)	74.36 (1.74)	74.47* (0.82)	74.70* (1.02)	**76.85** (1.96)
	F1-s	GloVe	**65.00** (1.36)	63.19 (2.32)	63.07 (3.51)	62.60* (2.74)	59.83* (2.46)	61.51 (2.66)	61.90 (3.32)
		TL	66.92 (2.41)	67.76 (1.75)	67.26 (2.27)	66.00 (3.03)	66.92* (1.21)	66.63* (1.61)	**69.91** (2.72)
	extraction	GloVe	76.00* (0.99)	75.98* (1.17)	**77.99** (1.14)	76.43 (1.57)	75.81 (1.57)	77.81 (1.10)	76.76 (1.69)
		TL	83.51 (1.09)	83.32 (0.94)	83.88 (0.88)	**84.21** (1.81)	83.29 (1.15)	83.48 (1.30)	82.90 (0.72)
	targeted	GloVe	54.65 (1.37)	53.67 (0.94)	**54.85** (0.99)	53.73 (1.93)	51.65* (2.32)	54.17 (2.26)	54.37 (2.56)
		TL	62.89 (1.18)	62.29 (1.32)	62.55 (1.66)	62.61 (1.79)	62.03 (1.14)	62.35 (0.77)	**63.70** (1.14)
MPQA	acc-s	GloVe	**78.18** (2.72)	74.37 (3.47)	75.94 (3.02)	77.38 (4.91)	72.82* (3.88)	73.83* (2.30)	73.01* (3.41)
		TL	71.84 (3.46)	72.01 (3.01)	**73.08** (4.01)	70.96 (2.03)	72.79 (3.13)	72.61 (3.84)	70.47 (1.51)
	F1-s	GloVe	**42.03** (1.50)	39.96 (1.66)	40.58 (1.28)	41.16 (2.32)	39.19 (1.94)	39.90* (1.06)	39.00 (1.86)
		TL	39.92 (1.15)	40.27 (0.80)	**41.13** (3.05)	39.17 (0.79)	39.84 (1.56)	39.90 (1.66)	39.25 (0.68)
	extraction	GloVe	24.17 (1.44)	22.93* (1.57)	24.58 (1.36)	22.84* (1.58)	22.90* (2.49)	**25.34** (0.46)	24.77 (3.55)
		TL	30.98* (2.40)	30.71* (1.10)	31.19* (2.69)	31.41* (1.16)	31.41* (0.51)	31.88* (2.62)	**34.99** (1.10)
	targeted	GloVe	**18.88** (1.17)	17.03* (1.12)	18.66 (1.22)	17.60* (0.57)	16.70* (2.26)	18.70 (0.25)	18.11 (2.83)
		TL	22.25* (2.00)	22.09* (0.70)	22.74 (1.68)	22.30* (1.19)	22.86* (0.98)	23.05* (0.88)	**24.66** (1.07)

Table 9: *acc-s*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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 TL at a 95% confidence level.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
MAMS	acc-s	GloVe	**81.72** (1.12)	81.42 (0.32)	81.24 (0.44)	81.10 (0.46)	80.90 (0.70)	81.58 (0.96)	81.70 (0.79)
		TL	84.59 (0.50)	**84.81** (0.78)	83.99* (0.84)	83.73 (0.68)	84.28 (0.39)	83.90* (0.51)	84.67 (0.56)
	F1-s	GloVe	**81.05** (1.15)	80.71 (0.43)	80.53 (0.60)	80.39 (0.69)	80.22 (0.80)	80.81 (1.01)	80.94 (0.88)
		TL	84.24 (0.51)	**84.39** (0.80)	83.59* (0.87)	83.30 (0.67)	83.89 (0.32)	83.44* (0.60)	84.24 (0.58)
	extraction	GloVe	76.49 (0.99)	76.21* (0.39)	76.49 (1.06)	76.87 (0.64)	76.83 (0.48)	76.97 (0.54)	**77.36** (0.19)
		TL	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	GloVe	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)
		TL	65.17 (0.35)	65.10 (0.63)	64.65* (0.88)	65.00* (0.48)	64.50* (0.79)	64.88 (0.46)	**65.70** (0.55)
Restaurant	acc-s	GloVe	83.02 (1.82)	83.23 (1.69)	83.26 (0.89)	**83.80** (0.78)	83.01 (1.16)	83.36 (1.09)	83.65 (0.48)
		TL	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	GloVe	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)
		TL	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	GloVe	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)
		TL	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	GloVe	65.06 (2.66)	**66.06** (2.63)	65.89 (1.32)	65.82 (1.31)	65.16 (1.50)	65.73 (1.46)	65.60 (1.06)
		TL	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*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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 TL at a 95% confidence level.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop	acc-s	GloVe	77.47 (1.43)	77.53 (0.69)	76.52 (1.38)	**78.34** (2.23)	77.32 (1.06)	76.67 (1.77)	77.22 (1.05)
		TL	81.22 (1.42)	79.63* (1.56)	80.54 (2.27)	80.05* (1.28)	80.87 (1.64)	79.31* (1.12)	**81.89** (1.07)
	F1-s	GloVe	71.55 (0.81)	70.75 (2.75)	70.61 (2.46)	**72.04** (3.85)	70.70 (1.83)	67.82 (1.20)	67.27 (2.80)
		TL	74.53 (3.09)	74.84 (1.87)	74.89 (3.32)	73.70 (1.98)	75.22 (2.03)	71.90* (1.77)	**75.33** (1.61)
	extraction	GloVe	74.87 (1.15)	73.55 (2.09)	**75.70** (0.96)	74.81 (1.18)	74.75 (0.57)	74.15* (1.84)	75.06 (1.06)
		TL	80.82 (1.35)	80.89 (0.83)	80.39 (1.02)	**81.86** (1.24)	80.05* (0.59)	81.37 (0.99)	81.23 (0.82)
	targeted	GloVe	57.99 (0.69)	57.02 (1.53)	57.92 (1.08)	**58.62** (2.19)	57.80 (1.10)	56.82 (0.71)	57.97 (1.24)
		TL	65.62* (0.76)	64.42 (1.64)	64.73 (1.30)	65.52* (0.52)	64.72* (0.97)	64.55* (1.53)	**66.51** (0.43)
MPQA	acc-s	GloVe	87.75 (3.15)	88.65 (4.20)	87.64 (3.71)	**89.11** (3.29)	86.85 (1.46)	88.16 (3.20)	85.29* (0.89)
		TL	88.63 (1.83)	**90.08** (2.94)	87.23 (1.80)	85.62 (4.14)	88.71 (2.89)	88.75 (1.56)	88.01 (1.36)
	F1-s	GloVe	54.18 (8.15)	**59.87** (14.46)	51.51 (8.83)	56.39 (10.28)	48.09 (4.83)	52.32 (9.96)	45.92 (3.06)
		TL	52.83* (3.40)	55.80 (5.28)	**59.03** (5.99)	54.48 (9.33)	56.55 (3.28)	53.82 (7.01)	55.74 (6.52)
	extraction	GloVe	20.68 (0.65)	21.00 (1.73)	20.78 (1.52)	20.48* (0.85)	19.54* (2.18)	**21.73** (0.74)	20.11 (2.38)
		TL	32.33 (3.71)	30.39* (1.09)	31.75 (1.78)	32.18 (1.29)	30.65* (1.52)	31.00* (1.92)	**33.38** (0.67)
	targeted	GloVe	18.14 (0.58)	18.57 (1.13)	18.18 (1.10)	18.23 (0.55)	16.94* (1.66)	**19.16** (0.89)	17.16 (2.07)
		TL	28.60 (2.82)	27.35* (0.72)	27.67* (1.29)	27.53* (1.30)	27.15* (0.92)	27.49* (1.38)	**29.39** (0.95)

Table 11: *acc-s*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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 TL at a 95% confidence level.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
MAMS	acc-s	GloVe	80.73* (0.22)	80.98 (1.17)	80.89* (0.45)	80.68* (0.48)	80.87* (0.53)	81.36 (0.82)	**82.14** (0.77)
		TL	84.37 (0.22)	83.99* (0.25)	**84.73** (0.34)	84.18 (0.49)	84.17 (0.53)	83.79* (0.40)	84.43 (0.41)
	F1-s	GloVe	80.20* (0.24)	80.48* (1.15)	80.38* (0.50)	80.14* (0.57)	80.38* (0.56)	80.76 (0.95)	**81.67** (0.77)
		TL	84.14* (0.24)	83.72* (0.21)	**84.46** (0.35)	83.96* (0.47)	83.93 (0.47)	83.54* (0.42)	84.12 (0.45)
	extraction	GloVe	78.93 (0.66)	79.11 (0.52)	**79.24** (0.62)	79.00 (0.47)	78.86 (0.67)	78.68 (0.35)	79.15 (0.39)
		TL	77.81 (0.48)	77.86 (0.59)	77.62* (0.34)	78.32 (0.31)	77.59* (0.21)	**78.54** (0.38)	78.35 (0.40)
	targeted	GloVe	63.72* (0.63)	64.06* (0.66)	64.10* (0.48)	63.74* (0.28)	63.76* (0.21)	64.01* (0.62)	**65.01** (0.44)
		TL	65.65 (0.39)	65.39* (0.47)	65.77* (0.44)	65.93 (0.20)	65.31* (0.56)	65.81 (0.41)	**66.15** (0.54)
Restaurant	acc-s	GloVe	78.42* (0.78)	78.78 (0.67)	**79.58** (0.89)	78.75 (0.52)	78.31* (0.78)	79.14 (0.41)	78.76 (0.37)
		TL	81.90 (0.69)	**81.90** (0.69)	81.53 (0.86)	81.89 (0.84)	81.02 (0.85)	80.47* (1.10)	81.77 (0.46)
	F1-s	GloVe	62.89 (2.84)	64.01 (2.56)	**65.37** (1.47)	62.49* (0.86)	62.54* (2.28)	63.00* (1.64)	63.15* (1.94)
		TL	67.98 (3.46)	69.26 (1.07)	69.09 (1.72)	68.18 (1.90)	67.54 (3.88)	67.14* (1.05)	**69.37** (0.97)
	extraction	GloVe	78.22* (0.78)	79.20 (1.07)	78.85 (1.08)	78.38* (0.73)	78.21 (1.37)	**79.62** (0.48)	79.18 (0.76)
		TL	81.69 (0.71)	81.84 (0.88)	**82.56** (0.79)	82.25 (0.22)	82.07 (0.68)	82.48 (0.61)	82.33 (0.52)
	targeted	GloVe	61.34* (0.73)	62.39 (0.95)	62.75 (1.26)	61.73* (0.77)	61.25* (1.26)	**63.02** (0.42)	62.36 (0.60)
		TL	66.90 (0.40)	67.03 (1.06)	67.31 (0.32)	**67.36** (0.66)	66.49 (0.52)	66.38* (1.31)	67.32 (0.55)

Table 12: *acc-s*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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.

D Additional Negation and Speculation Result Tables

		NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
					Laptop$_{Neg}$			
F1-s	GloVe	40.88 (2.17)	38.11* (1.71)	38.89 (3.32)	**42.33** (1.19)	39.46* (2.98)	38.11* (2.02)	37.13* (2.78)
	TL	44.81 (2.40)	45.05 (3.47)	44.58 (2.29)	43.53 (2.74)	43.83 (1.90)	44.77 (1.54)	**45.08** (2.68)
					Restaurant$_{Neg}$			
F1-s	GloVe	46.58 (3.24)	44.16* (2.18)	42.74* (1.09)	**47.65** (1.35)	44.00 (3.99)	44.78 (1.85)	44.14* (1.89)
	TL	52.85* (1.69)	54.41 (1.51)	54.08 (3.87)	52.54* (1.99)	**55.63** (1.65)	52.16* (2.01)	53.59* (1.95)

Table 13: F_1-*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.

		NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
					Laptop$_{Spec}$			
F1-s	GloVe	32.74 (2.35)	33.14 (0.98)	**35.24** (2.16)	33.62 (2.59)	33.83 (1.62)	33.21 (1.00)	31.99* (1.92)
	TL	33.02 (4.07)	33.33 (1.56)	33.14 (2.10)	31.72* (0.88)	33.25 (2.14)	32.71 (1.38)	**34.08** (1.40)
					Restaurant$_{Spec}$			
F1-s	GloVe	55.27 (3.82)	**57.77** (2.91)	56.27* (2.36)	55.59* (1.09)	57.35 (3.75)	56.55 (3.14)	57.32 (2.30)
	TL	58.84* (1.58)	60.95 (0.98)	**62.36** (2.86)	58.44* (2.24)	60.52 (3.25)	59.23 (1.81)	60.74 (2.32)

Table 14: F_1-*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.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop$_{Neg}$	acc-s	GloVe	**37.07** (3.78)	32.99* (0.96)	33.98 (2.60)	36.64 (3.17)	34.83 (1.09)	32.09* (3.41)	29.45* (1.46)
		TL	41.84* (0.54)	39.38* (2.69)	40.85* (2.50)	**45.40** (1.70)	43.00* (1.57)	38.77* (2.64)	41.15* (3.28)
	F1-s	GloVe	34.02 (2.05)	31.48* (2.66)	33.06 (2.21)	**34.61** (3.34)	33.66 (1.13)	28.92* (4.11)	24.92* (1.33)
		TL	38.26* (1.67)	38.05 (3.06)	39.42 (4.02)	**42.32** (3.45)	41.39 (1.97)	36.73 (3.67)	38.92 (3.45)
	extraction	GloVe	72.49 (1.45)	74.40 (1.64)	73.91 (0.99)	71.25* (1.20)	73.83 (0.93)	73.33 (1.90)	**74.60** (1.51)
		TL	80.94 (1.64)	81.31 (1.29)	81.21 (1.36)	81.94 (1.46)	79.00* (0.87)	81.92 (1.36)	**82.75** (1.80)
	targeted	GloVe	**26.87** (2.75)	24.56 (1.21)	25.12 (1.99)	26.14 (2.55)	25.71 (0.68)	23.54 (2.66)	21.98* (1.26)
		TL	33.86* (0.71)	32.04* (2.54)	33.14* (1.56)	**37.20** (1.59)	33.98* (1.47)	31.78* (2.46)	34.01* (2.29)
Restaurant$_{Neg}$	acc-s	GloVe	46.02 (4.88)	43.13* (3.24)	41.06* (3.36)	**49.02** (1.31)	41.65* (4.06)	44.02* (3.09)	42.69* (2.01)
		TL	53.79 (2.56)	54.40 (3.63)	52.25 (3.63)	**54.42** (3.18)	53.16 (1.92)	54.31 (2.49)	52.25 (2.51)
	F1-s	GloVe	40.03 (5.30)	38.56* (3.10)	37.54 (3.69)	**41.05** (2.45)	37.01 (4.16)	38.03 (3.40)	38.45 (2.28)
		TL	48.03 (3.48)	49.13 (2.62)	48.31 (3.71)	49.18 (3.58)	48.80 (2.03)	**49.42** (2.76)	48.06 (2.12)
	extraction	GloVe	81.74 (0.77)	82.37 (0.64)	82.36 (0.80)	80.61* (0.72)	81.34 (1.48)	**82.38** (1.00)	81.32 (0.37)
		TL	84.08 (0.72)	82.87* (0.66)	84.32 (0.68)	83.71 (0.55)	83.45 (1.09)	84.02 (1.12)	**84.84** (0.99)
	targeted	GloVe	37.61 (3.86)	35.54* (2.81)	33.82* (2.78)	**39.51** (1.03)	33.88* (3.38)	36.27* (2.64)	34.71* (1.58)
		TL	45.23 (2.19)	45.07 (2.90)	44.04 (2.77)	45.57 (2.91)	44.38 (2.05)	**45.62** (1.88)	44.31 (1.74)

Table 15: *acc-s*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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.

			NEG$_{CD}$	DR	LEX	NEG$_{SFU}$	SPEC	UPOS	STL
Laptop$_{Spec}$	acc-s	GloVe	33.56* (2.39)	35.00 (2.94)	31.70* (1.64)	34.67 (2.21)	**37.24** (2.48)	31.59* (1.57)	32.57* (1.94)
		TL	34.73 (1.94)	35.02 (3.50)	**35.16** (2.48)	32.79 (1.43)	34.24 (1.95)	34.27 (1.56)	34.28 (1.19)
	F1-s	GloVe	32.66 (2.33)	33.99 (3.14)	30.72* (1.33)	33.20 (2.36)	**34.99** (2.91)	29.88* (1.50)	30.75 (2.02)
		TL	33.60 (2.30)	**34.14** (3.47)	33.83 (2.83)	30.92 (1.48)	33.36 (1.98)	32.45 (1.32)	32.47 (1.67)
	extraction	GloVe	71.10 (1.97)	69.42 (2.25)	72.61 (0.95)	70.46 (1.41)	69.33* (1.58)	70.58* (2.46)	**73.04** (2.48)
		TL	80.50 (0.95)	80.25 (0.90)	79.91* (0.61)	80.83 (0.61)	79.35* (1.34)	80.95 (0.43)	**82.00** (1.32)
	targeted	GloVe	23.88 (2.05)	24.26 (1.81)	23.02* (1.35)	24.44* (1.81)	**25.82** (1.82)	22.32* (1.73)	23.80 (1.77)
		TL	27.96 (1.60)	28.11 (2.92)	28.08 (1.81)	26.51 (1.34)	27.16 (1.39)	27.74 (1.31)	**28.12** (1.32)
Restaurant$_{Spec}$	acc-s	GloVe	35.54* (0.90)	**40.09** (2.95)	37.98 (2.38)	37.18 (2.14)	37.97 (2.28)	38.32 (1.31)	37.23 (1.65)
		TL	38.80 (2.17)	38.50 (1.19)	40.72 (1.18)	**40.84** (2.30)	39.69 (1.69)	39.49 (0.89)	40.55 (1.16)
	F1-s	GloVe	31.46* (1.47)	**35.99** (3.98)	34.37 (2.89)	32.53 (1.82)	34.02 (2.03)	32.92 (1.74)	33.18 (1.15)
		TL	33.00* (3.34)	33.28 (1.65)	**35.92** (1.84)	35.26 (3.37)	34.47 (3.78)	35.08 (1.78)	35.40 (1.59)
	extraction	GloVe	78.26* (0.92)	80.38 (1.86)	80.33 (1.13)	79.75* (1.05)	79.99* (1.20)	**81.98** (1.04)	80.59 (1.29)
		TL	83.91 (0.67)	83.93 (1.03)	84.60 (0.63)	84.77 (0.29)	83.40 (1.58)	**85.03** (1.47)	84.85 (0.83)
	targeted	GloVe	27.82* (0.79)	**32.25** (2.79)	30.51* (2.06)	29.64* (1.60)	30.37 (1.91)	31.43 (1.42)	30.02 (1.74)
		TL	32.56 (1.83)	32.31* (1.13)	34.45 (1.00)	**34.62** (1.96)	33.08 (1.01)	33.59 (1.20)	34.41 (1.23)

Table 16: *acc-s*, F_1-*s*, extraction (F_1-*a*) and full targeted (F_1-*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.

E Hyperparameter Search Space

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	0.6156
Training duration	14232 sec
Model implementation	https://bit.ly/3lAz6yf

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/3lAz6yf

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 (NEG$_{CD}$) dataset.

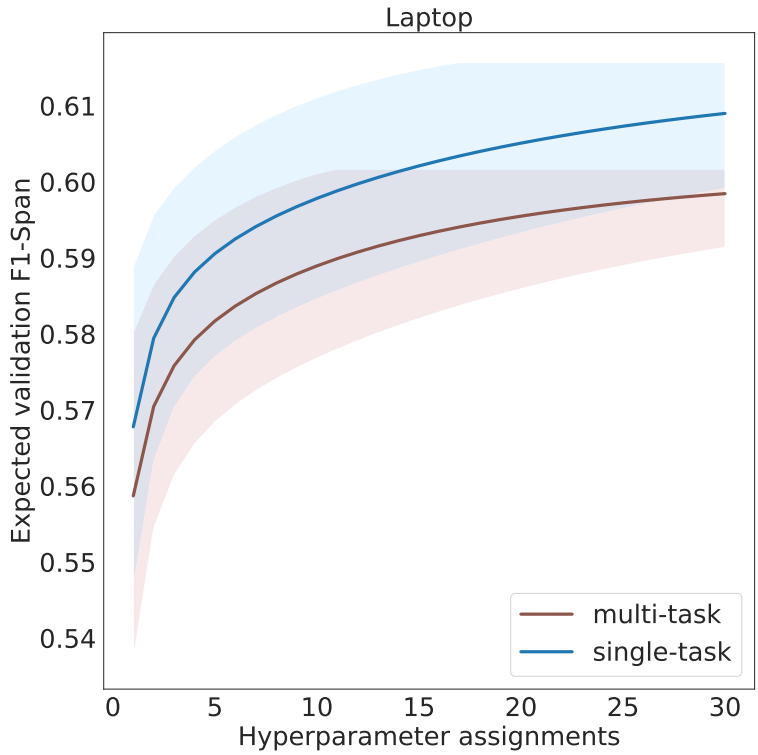

Figure 2: Hyperparameter budget against expected span F1/F_1-i performance for the STL and MTL models. The hyperparameter search space is stated within Tables 17 and 18 for the STL and MTL models respectively. The shaded areas represent the expected performance ± 1 standard deviation. Note the shaded area does not go beyond the maximum observed validation score as recommended Dodge et al. (2019).

F Additional Reproducibility Statistics

Embedding	Model	Number of Parameters			
		Including Auxiliary Task			
		Yes		No	
		Trainable	All	Trainable	All
GloVe	STL	364,042	1,785,967	364,042	1,785,967
	NEG_{CD}	385,851	2,403,876	385,122	2,403,147
	NEG_{SFU}	385,851	7,066,176	385,122	7,065,447
	SPEC	385,851	7,066,176	385,122	7,065,447
	UPOS	388,413	2,952,738	385,122	2,949,447
	DR	397,213	2,961,538	385,122	2,949,447
	LEX	1,191,204	3,755,529	385,122	2,949,447
TL	STL	1,001,170	56,870,931	1,001,170	56,870,931
	NEG_{CD}	1,051,939	56,921,700	1,051,210	56,920,971
	NEG_{SFU}	1,051,939	56,921,700	1,051,210	56,920,971
	SPEC	1,051,939	56,921,700	1,051,210	56,920,971
	UPOS	1,054,501	56,924,262	1,051,210	56,920,971
	DR	1,063,301	56,933,062	1,051,210	56,920,971
	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. The auxiliary task specific layer is shown 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. There are many more trainable parameters for the MTL models ignoring the auxiliary task parameters. This is because the hyperparameter search finds a larger shared LSTM hidden dimension to be preferable for the MTL models (see Tables 17 and 18). For the GloVe MTL models the total number of parameters changes depending on the auxiliary task. This is because the GloVe embeddings contain different numbers of vocabulary words, as we filter words based on those in the auxiliary and main task datasets/corpora. The large difference in the number of trainable parameters between GloVe and TL models is due to the fact that the TL is 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)
GloVe	STL	CPU	1	10.24	10.45
			8	7.00	7.21
			16	6.67	6.91
			32	6.35	6.51
		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
TL	STL	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
	MTL	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 TL 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 (NEG_{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.