Attention-with-logits: an End-to-end Weakly Supervised Approach for Sentence-level Sentiment Classification

ABSTRACT

Existing methods for sentence-level sentiment classification are mainly based on supervised learning models, that require large datasets of labelled sentences to train the classifiers. Manually labelling textual data at sentence level is very expensive. To reduce the labelling efforts, one can use datasets labelled at document level to train the classifiers based on weakly supervised approaches. The simplest weakly-supervised approach is treating each document as a single sentence, and using generic end-to-end supervised learning models, such as SVM, neural network works, etc., to train the classifiers. Though it is obviously problematic as it loses the sentence-level sentiment signals, it has been widely adopted in many applications thanks to its ease in implementation. In this paper, we propose a novel approach, called 'attention-wtih-logits' (AWL), that can also easily adapt a generic end-to-end supervised model into a weakly-supervised model for sentence-level sentiment classification, but without ignoring the sentence-level sentiment signals. In this approach, we first feed unlabelled sentences into a generic supervised model, and then combine the resulting sentence-level logit vectors with an attention mechanism into document-level logit vectors. The document-level logit vectors and the document-level labels are used to define the loss of the supervised model. By minimising the document-level loss, the parameters of the sentence-level model can be learned. Our experiment results show that, the proposed approach outperforms the documents-as-sentences approach, by an average margin of 6% .

CCS CONCEPTS

• **Computing methodologies → Information extraction**; *Natural language processing*.

KEYWORDS

sentence-level sentiment classification, weak supervision, review data, neural network models

1 INTRODUCTION

Supervised machine learning models have been widely applied in sentiment classification [3, 5, 10, 11, 16, 18, 20]. To train such classifiers, one needs to provide training data labelled at the same granularity as that of the sentiment classification task at hand. For example, it requires labelled documents to train document-level sentiment classifiers, and labelled sentences sentence-level sentiment classifiers. To label the data, one needs to manually read the textual content of each training example, and decides which sentiment class each sample is associated with. Therefore, the labelling process usually takes laborious manual efforts. Also, the finer the granularity, the more difficulties to make the labelling decisions and therefore more manual efforts are needed to prepare the data. In practice, datasets of labelled sentences are much more expensive than datasets of labelled documents to obtain. The difficulty in the sentence labelling process can be reflected by the fact that only few datasets containing sentence-level sentiment labels are publicly available on the internet. The unavailability of labelled sentence datasets limits the feasibility of training effective sentence-level sentiment classifiers, especially in today's era when many methods are built upon deep learning models that usually require very large datasets for the training processes.

One possible way to ease the problem is using labelled documents that are less expensive, sometimes even almost free (such as on-line reviews with ratings) to obtain, to train sentence-level sentiment classifiers. The easiest way to do that is treating each document as a single long sentence, and applying generic methods for sentence-level sentiment classification to train the classifiers.

However, there exists a problem in this documents-as-sentences approach: the training process is guided only by the document-level sentiments and loses the sentiment signals of individual sentences. There exists a high possibility that the sentiments of individual sentences may be very different from the sentiments of their containing documents. For example, there may exist very negative sentences in an overall positive document, or positive sentences in an overall negative document. By using document-level sentiments as the supervision signal, the true sentiment orientations of individual sentences can be easily misrepresented, that would inevitably result in harmful impacts to the classifiers. Furthermore, sequential deep learning models, such as GRU and LSTM, cannot handle long word sequences. To train classifiers with these models on long documents, one has to either shorten the documents by throwing away words or use the truncated back-propagation technique [1] in the training process. The former way leads to loss of information, the latter way may significantly increase the training time.

This paper proposes a novel end-to-end approach, called 'attention-with-logits' (AWL), that uses labelled documents to train sentence-level sentiment classifiers. Unlike the documents-as-sentences approach, the proposed approach feeds individual sentences into end-to-end supervised learning models, and combines the resulting logit vectors of sentences from the same reviews by an attention mechanism into document-level logit vectors. The cross entropies between the resulting document-level logit vectors and the document-level labels are used as the loss of the supervised learning models. The parameters of the supervised learning models and the attention mechanism will be learned simultaneously by minimising the loss. In the experiment, we collected two review datasets, one containing 50,000 hotel reviews crawled from TripAdvisor, the other 50,000 reviews crawled from Amazon. We used the review data to train sentence-level sentiment classifiers and the results indicate that, the proposed approach can effectively improve the performance of the weakly-supervised classifiers.

The rest of this paper is organised as follows: Section 2 presents a brief introduction to the related work; Section 3 introduces the proposed approach; Section 4 presents the evaluation results and Section 5 provides the conclusion and future research direction.

2 RELATED WORK

We divide existing methods for sentence-level sentiment classification into 3 categories: lexicon-based methods; end-to-end supervised machine learning methods; weakly-supervised machine learning methods.

Lexicon-based models require lexicons consisting of opinion words or phrases. These methods assume that opinion words and phrases are the dominating indicator for sentiment classification, therefore, the input features of those methods are usually derived based on the presence or absence of the opinion words in each sentence. A big array of models, ranging from rule-based algorithms[4, 15, 25], to unsupervised [13, 14, 32] and supervised machine learning models [2, 7, 19, 24, 30], have been proposed to build sentiment classifiers with the help of sentiment lexicons.

End-to-end supervised machine learning methods usually rely only on the statistical pattern of the training corpuses to learn the classifiers. These methods are very susceptible to how the input sentences are represented. Bag-of-words (BOW) representation models, such as one-hot encoding, tf-idf, topic models [22], had been widely used in the early days of sentiment analysis. BOW-based models usually suffer from such a problem: they are inefficient in representing contextual and semantic information. In recent years, distributed embedding representations, such as W2V [17], Para2Vec [12], Skip-thought [9], BERT [27], that are usually built upon generative languages models to encode the contextual and semantical information, have been proven more effective in sentence-level sentiment classification. To compensate for the lack of sentiment lexicons, methods in this category usually use complex learning architectures to extract high-level discriminative features. This is especially true recently as many newly developed methods are centred around an array of deep learning models, such as LSTM, GRU, CNN, Transformer [8, 23, 27, 29], etc.

In weakly-supervised machine learning category, besides the documents-as-sentences approach, there are also existing a few methods that require external knowledge, such as lexicons, handcrafted classification rules, sentence-level labels, etc., to train the classifiers. Qu et al. [21] proposed a multi-expert model that makes use of document-level sentiments and opinion lexicons to build an ensemble of base classifiers. The sentiment of an unseen sentence is decided by the votes of the base classifiers. Yang et al. [33] proposed a CRF model that uses sentence-level sentiment labels as the main supervision signal, and uses overall ratings as a form of posterior regularisation to keep sentence-level sentiments and document-level sentiments consistent. Tackstrom et al. [26] proposed a hidden CRF model, that treats the sentence-level sentiments as latent variables, and the overall ratings as observable variables conditioning on the latent variables. Opinion lexicons are used to define the feature functions for the CRF model. Wu et al. [31] proposed the SSWS model that uses two levels of features: document-level sentiments and word-level sentiments, along with a predefined set of linguistic rules, to train a linear sentence-level sentiment classifier.

In the research, we differ our research goal from these weakly-supervised models by focusing only on generalised, end-to-end approaches to train the sentence-level sentiment classifiers, to avoid the laborious efforts to build the external knowledge resources.

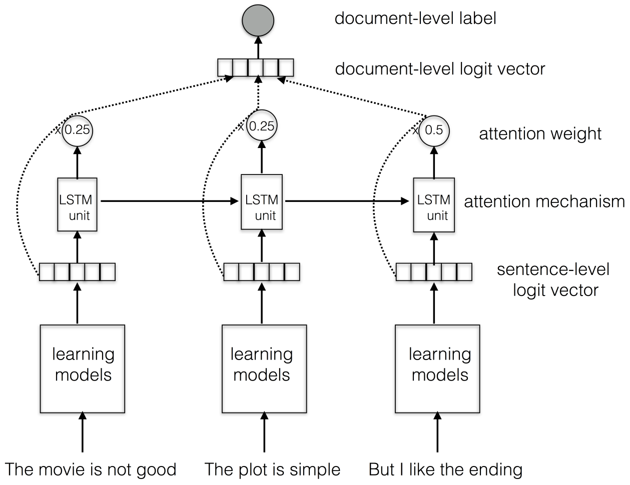

Figure 1: The structure of the AWL approach. In the proposed approach, we first feed each sentence of a document into a supervised neural network model, and then combine the resulting sentence-level logit vectors into the document-level logit vector in order to associate sentence-level sentiments with document-level sentiments. The LSTM model is used to compute the attention weight of each sentence in the combination process.

3 THE ATTENTION-WITH-LOGITS MODEL

Assuming a neural network model is used to train the classifier. Given a document d consisting of N sentences, each sentence is fed into the neural network. Each sentence will result in a logit vector in the output layer of the neural network model, that is usually used to computed the probability distribution of the sentence over all the sentiment classes. If the sentence-level ground-truth labels are available, the cross entropies between the labels and the probability distributions can be used as the loss of the model.

Since we wish to avoid the laborious sentence labelling efforts, the sentence-level labels are not available. Instead, we assume only document-level labels are available and use them as the supervision signal to train the sentence-level classifier. To achieve that, we have to define a way to associate the sentence-level logit vectors with document-level labels. In the proposed approach, we combine the sentence-level logit vectors into document-level logit vectors, and use the cross-entropies between the document-level logit vectors and ground-truth labels as the loss. Therefore, a central piece in the proposed approach is the combination function to aggregate sentence-level logit vectors into document-level logit vectors.

In this paper, the sentence-level logit vectors are combined by an attention mechanism. Since the order of sentences may have impacts on the combined sentiments, we use a small LSTM model, to compute the attention weight of each sentence-level logit vector in the combination process.

Assuming the resulting logit vector of a sentence n of a review d in a neural network model is $z_d^{(n)}$. Assuming the mapping function of the LSTM attention unit is A, that maps the sentence-level logit vector to its hidden state vector $h_d^{(n)}$:

$$h_d^{(n)} = A(z_d^{(n)}) \tag{1}$$

The hidden state is further fed into a dense layer with an output size of 1 to get the attention score:

$$s_d^{(n)} = v^{\top} h_d^{(n)} \tag{2}$$

where v is the weight vector of the dense layer. The attention score is fed into the softmax function to compute the attention weight:

$$w_d^{(n)} = \frac{\exp\,(s_d^{(n)})}{\sum_{j=1}^{N} \exp\,(s_d^{(j)})} \tag{3}$$

The document-level logit vector of the document is computed as follows:

$$z_d = \sum_{j=1}^{N} w_d^{(j)} z_d^{(j)} \tag{4}$$

The cross entropies between the resulting document-level logit vectors and document-level ground truth labels are used as the loss of the proposed approach. The structure of the proposed approach is summarised in Figure 1.

In the test phase, given an unseen sentence m, we feed it to the learning model and simply use the resulting sentence-level logit z_m to predict its sentiment by:

$$s_m = \underset{k}{\operatorname{argmax}}\, z_{mk} \tag{5}$$

4 EVALUATION

This section presents the evaluation results. We collected two datasets of reviews, one containing 50,000 electronic product reviews from Amazon [6], the other 50,000 hotel reviews from TripAdvisor[28]. Each review consists a piece of text, and an overall rating on a 5-point scale that can serve as the document-level sentiment label. Statistics of the two datasets are shown in Table1.

Table 1: Statistics of the review datasets

	Num of reviews	Avg num of sentences per review	Avg num of words per review	Num of sentiment classes
TripAdvisor	50,000	9	176	5
Amazon	50,000	12	126	5

We also collected two sentence datasets, each containing 7000 labelled sentences extracted from reviews on the two sites. Each sentiment label takes 3 possible categorical values: positive, negative, neutral. Statistics of the labelled sentence sets are shown in Table 2. We use the labelled review sentences to train sentence-level sentiment classifiers based on a number of supervised models. To train the classifiers, 5,000 labelled sentences are randomly chosen from each dataset as the training set, and the remaining as the test set.

Table 2: Statistics of the review sentence sets

	Training size	Test size	Avg num of words	Num of classes
TripAdvisor	5000	2000	22	3
Amazon	5000	2000	17	3

At the meantime, we use the reviews to train the sentence-level sentiment classifiers based on the proposed approach and the documents-as-sentences approach. The performance of those classifiers is also evaluated on the test sets of the labelled sentences. It is noteworthy that, since the ratings are on a 5-star scale, the predictions of the classifiers trained under the supervision of ratings would also be values on that scale. Therefore, the predicted labels have to be converted into the same 'negative-neutral-positive' format used by labelled sentences in the evaluation process. In this paper, the following conversion scheme is used: 'negative' when predictions are less than 3 stars, 'neutral' when equal to 3 stars, 'positive' otherwise. The performance comparison among all the classifiers will be shown in Section 4.2.

Naturally, the performance of the classifiers trained with the proposed approach should be lower than that of the classifiers trained with generic supervised models when their training sizes are close [1]. However, by increasing the training size of the proposed approach simply by adding more raw reviews, the performance gap is expected to narrow down. Details of the effects of training size on the performance of the classifiers will be shown in Section 4.3.

It is also possible to mix raw reviews with labelled sentences to train the classifiers with the proposed approach. In this case, each labelled sentence is treated as a review consisting of only one sentence. To mix the two types of data, we convert the ratings of the reviews into the negative-neutral-positive format with the same scheme mentioned previously. The performance of the classifiers trained on the mixed data will be shown in Section 4.4.

4.1 Pre-processing and Hyper-parameter Setting

In the experiment, other than sentence tokenisation and removing the words that appear for less than 5 times in the datasets, no additional pre-processing treatment is performed. Google's W2V model [2] pre-trained on the Google news corpus is used to represent words. The dimensionality of the W2V vectors is 300. In the training process, the W2V vectors are set as non-trainable. The Keras package with Tensorflow backend [3] is used to implement both the baseline models and the proposed approach.

[1] When labelled sentences are used to train the classifiers, each sentence is counted as a training sample; when reviews are used as the training data, each review is counted as a training sample

[2] https://code.google.com/archive/p/word2vec/

[3] https://www.keras.io/

4.2 Performance Comparison of Attention-with-logits and the baselines

As mentioned previously, we train 3 types of classifiers: classifiers trained with generic supervised machine learning models; classifiers trained with the documents-as-sentences approach; classifiers trained with the proposed AWL approach. To make a fair comparison, 5,000 reviews are only drawn from each review dataset for the weakly-supervised classifiers to make the training sizes of the 3 types of classifiers approximately equal. The following models: SVM, MLP, CNN, GRU, LSTM are used to train the first and second types of classifiers; MLP, CNN, GRU, LSTM are used to train the AWL classifiers. The hyper-parameters of each model is decided by the grid-search technique on each dataset. Details of the hyper-parameters will be shown in the Appendix section.

The performance comparison in terms of accuracy among all the classifiers is shown in Table 3. We first compare the performance of the generic supervised classifiers and the AWL classifiers. As the results indicate, the performance of the AWL classifiers is close to that of the supervised SVM classifier, and around 6-10% lower than that of the supervised neural network classifiers.

We then compare the performance of the AWL classifiers and the documents-as-sentences classifiers. As the results show that, the AWL classifiers enjoy good margins of 6%-8% on the Amazon dataset, and 3-5% on the TripAdvisor dataset. The advantage of the AWL classifiers is more obvious on the Amazon dataset. One possible reason is that there are more cases where sentence-level sentiments are inconsistent with document-level sentiments in the Amazon dataset.

4.3 Impact of training size on classification performance

One can simply use larger training datasets to improve the classifiers' performance. However, for the generic supervised classifiers, increasing the training sizes comes at the heavy cost of labelling many more review sentences, therefore, is out of the question in this experiment. The training sizes of the weakly supervised classifiers can be increased simply by adding more reviews into the original training sets. In the experiment, the remaining raw reviews are divided into 5 portions. We iteratively increase the training sizes of the classifiers by adding one portion of the reviews at a time into the training set. The performance of the classifies with varying training sizes is shown in Figure 2 .

The results show that, as the training size increases, the performance of all the classifiers increases. However, the performance of the AWL classifiers increases more obviously and quickly. On both datasets, when 3 or 4 portions of reviews are added into the training sets, the performance of all the AWL classifiers starts to overpass the best performance of the generic supervised classifiers trained on the 5,000 sentences; whereas the performance of the documents-as-sentences classifiers is still obviously lower than that of the generic classifiers even when all the 5 portions of reviews are added into the training datasets.

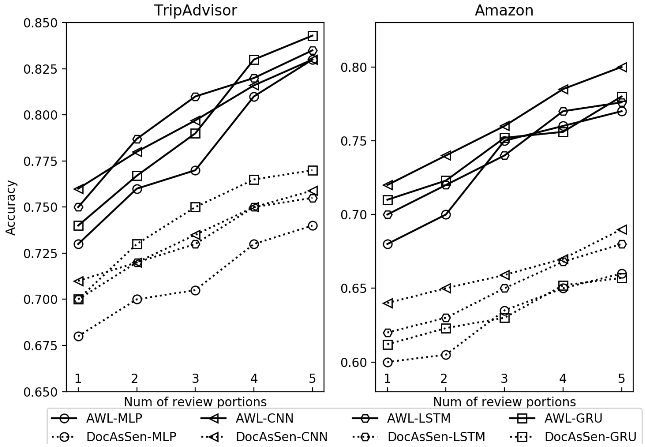

Figure 2: Impact of training size on the performance of the AWL classifiers and documents-as-sentences classifiers

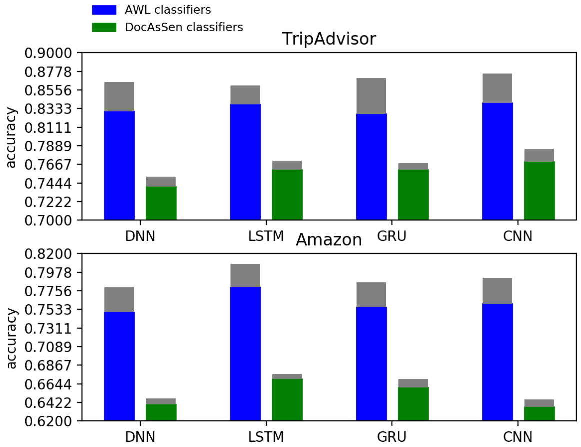

Figure 3: Performance improvement results from the addition of labelled sentences. The blue or green section of each bar represents the performance of the classifiers trained with 50,000 reviews, the grey section the improvement.

4.4 Mixing reviews and labelled sentences as the training data

It is also possible to mix a small number of labelled sentences with reviews to improve the performance of the weakly-supervised classifiers. We randomly draw 500 labelled sentences from each sentence dataset and mix them with all the 50,000 reviews of each review dataset, to train the classifiers with the proposed approach and the documents-as-sentences approach. The improvements result from the addition of the labelled sentences are shown in Figure 3.

As shown in the figure, the labelled sentences result in little change for the documents-as-sentences approach. The addition of the labelled sentences improves the performance of the AWL classifiers by an average margin of 2.9% on the Amazon dataset, 3.4% on the TripAdvisor dataset; whereas improves the performance of

Table 3: Performance of the 3 types of classifiers

Model	Approach	Training data type	Input features	TripAdvisor	Amazon
SVM	Supervised	sentences	ParaVec	0.71	0.66
	DocAsSen	reviews	ParaVec	0.58	0.49
MLP	Supervised	sentences	W2V mean	0.75	0.70
	DocAsSen	reviews	W2V mean	0.64	0.55
	AWL	reviews	W2V mean	0.68	0.63
CNN	Supervised	sentences	W2V	0.81	0.75
	DocAsSen	reviews	W2V	0.68	0.62
	AWL	reviews	W2V	0.73	0.69
LSTM	Supervised	sentences	W2V	0.77	0.73
	DocAsSen	reviews	W2V	0.66	0.58
	AWL	reviews	W2V	0.69	0.65
GRU	Supervised	sentences	W2V	0.80	0.72
	DocAsSen	reviews	W2V	0.66	0.60
	AWL	reviews	W2V	0.71	0.66

the generic classifiers by 0.8% on the Amazon dataset, and 1.2% on the TripAdvisor dataset. The improvements on the AWL classifiers are much more obvious. One possible reason is that, each training instance in the documents-as-sentences approach is a review, that contains many more words than the sentences, and therefore has dominantly stronger impacts in the training process. This makes it difficult for the models to capture the information of the added review sentences.

5 DISCUSSION AND CONCLUSION

In this paper we proposed a novel end-to-end weakly supervised approach, called ‘attention-with-logits’, to use labelled documents to train sentence-level sentiment classifiers. Experiment results show that, the proposed approach outperforms the widely used documents-as-sentences approach by a significant margin.

In the experiment, we used MLP, LSTM, GRU, and CNN to train the classifiers based on the proposed approach. It is noteworthy that the proposed approach is not only applicable on these models, but also applicable on any other learning models that output logit vectors for estimating the probability distribution of each training example over the target classes. It may also be possible to use the proposed approach for other classification tasks in natural language processing, such as product aspect classification, question intent classification, etc. In the near future, we will explore these possibilities to have better understanding of its application potential.

REFERENCES

[1] Mikael Boden. 2002. A guide to recurrent neural networks and backpropagation. *the Dallas project* (2002).

[2] Yejin Choi and Claire Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 793–801. http://www.aclweb.org/anthology/D08-1083

[3] Hang Cui. [n. d.]. Comparative Experiments on Sentiment Classification for Online Product Reviews. ([n. d.]), 6.

[4] Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A Holistic Lexicon-based Approach to Opinion Mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 231–240. https://doi.org/10.1145/1341531.1341561

[5] Michael Gamon. [n. d.]. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. ([n. d.]), 7.

[6] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 507–517.

[7] Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence* 22, 2 (2006), 110–125.

[8] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]* (Aug. 2014). http://arxiv.org/abs/1408.5882 arXiv: 1408.5882.

[9] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. *arXiv:1506.06726 [cs]* (June 2015). http://arxiv.org/abs/1506.06726 arXiv: 1506.06726.

[10] Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 11, 538-541 (2011), 164.

[11] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[12] Quoc Le and Tomas Mikolov. [n. d.]. Distributed Representations of Sentences and Documents. ([n. d.]), 9.

[13] Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment Analysis with Global Topics and Local Dependency.. In *AAAI*, Vol. 10. 1371–1376.

[14] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 375–384.

[15] Cheng-Yu Lu, Shian-Hua Lin, Jen-Chang Liu, Samuel Cruz-Lara, and Jen-Shin Hong. 2010. Automatic event-level textual emotion sensing using mutual action histogram between entities. *Expert Systems with Applications* 37, 2 (March 2010), 1643–1653. https://doi.org/10.1016/j.eswa.2009.06.099

[16] Justin Martineau and Tim Finin. 2009. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. *Icwsm* 9 (2009), 106.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]* (Oct. 2013). http://arxiv.org/abs/1310.4546 arXiv: 1310.4546.

[18] Tony Mullen and Nigel Collier. 2004. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of EMNLP 2004*, Dekang Lin and Dekai Wu (Eds.). Association for Computational Linguistics, Barcelona, Spain, 412–418.

[19] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, Sydney, Australia, 611–618. http://www.aclweb.org/anthology/P/P06/P06-2079

[20] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *arXiv:cs/0205070* (May 2002). http://arxiv.org/abs/cs/0205070 arXiv: cs/0205070.

[21] Lizhen Qu, Rainer Gemulla, and Gerhard Weikum. 2012. A Weakly Supervised Model for Sentence-Level Semantic Orientation Analysis with Multiple Experts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, 149–159. http://www.aclweb.org/anthology/D12-1014

[22] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[23] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. http://www.aclweb.org/anthology/D13-1170

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

[25] Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *arXiv:cs/0212032* (Dec. 2002). http://arxiv.org/abs/cs/0212032 arXiv: cs/0212032.

[26] Oscar TÃďckstrÃűm and Ryan McDonald. 2011. Discovering Fine-Grained Sentiment with Latent Variable Structured Prediction Models. In *Advances in Information Retrieval*, Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch (Eds.). Vol. 6611. Springer Berlin Heidelberg, Berlin, Heidelberg, 368–374. https://doi.org/10.1007/978-3-642-20161-5_37

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (June 2017). http://arxiv.org/abs/1706.03762 arXiv: 1706.03762.

[28] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 783–792.

[29] Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2428–2437. http://aclweb.org/anthology/C16-1229

[30] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics* 35, 3 (Sept. 2009), 399–433. https://doi.org/10.1162/coli.08-012-R1-06-90

[31] Fangzhao Wu, Jia Zhang, Zhigang Yuan, Sixing Wu, Yongfeng Huang, and Jun Yan. 2017. Sentence-level Sentiment Classification with Weak Supervision. ACM Press, 973–976. https://doi.org/10.1145/3077136.3080693

[32] Fu Xianghua, Liu Guo, Guo Yanyan, and Wang Zhiqiang. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems* 37 (2013), 186–195.

[33] Bishan Yang and Claire Cardie. 2014. Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 325–335. http://www.aclweb.org/anthology/P14-1031

A APPENDIX: HYPER PARAMETERS OF THE AWL CLASSIFIERS USED IN THE EXPERIMENT

A.1 Global hyper-parameters used across in all the 4 neural network models

- Maximum sentence length:100
- Maximum number of sentences per review: 20
- LSTM for the attention mechanism–hidden state size: 100; dropout rate:0.10

A.2 MLP

Architecture: input layer+1st hidden layer+2nd hidden layer +3rd hidden layer+ output layer

- the 1st hidden layer–200 units; activation:tanh; dropout rate : 0.20; L2 regularisation strength: 0.005
- the 2nd hidden layer– 100 units; activation:tanh; dropout rate: 0.20; L2 regularisation strength: 0.005
- the 3rd hidden layer–50 units; activation:tanh; dropout rate: 0.20; L2 regularisation strength: 0.005

A.3 GRU/ LSTM

Architecture: GRU/LSTM units+ 1st dense layer +output layer

- GRU units–hidden state size: 500; L2 kernel regularisation strength: 0.001; L2 recurrent regularisation strength : 0.001; dropout rate:0.1
- dense layer–50 units; activation:tanh; dropout rate:0.2

A.4 CNN

Architecture: CNN layer+ 1st dense layer +output layer

- CNN layer: 100 2-gram filters and 100 3-gram filters; activation function: relu; L2 kernel regularisation: 0.005
- 1st dense layer: 50 units; activation function: tanh; dropout rate: 0.25