Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park and **Pascale Fung**
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology, Clear Water Bay
[pxuab,eeandreamad,cwuak,jhpark,pascale]@ust.hk

Abstract

In this paper, we propose Emo2Vec which encodes emotional semantics into vectors. We train Emo2Vec by multi-task learning six different emotion-related tasks, including emotion/sentiment analysis, sarcasm classification, stress detection, abusive language classification, insult detection, and personality recognition. Our evaluation of Emo2Vec shows that it outperforms existing affect-related representations, such as Sentiment-Specific Word Embedding and DeepMoji embeddings with much smaller training corpora. When concatenated with GloVe, Emo2Vec achieves competitive performances to state-of-the-art results on several tasks using a simple logistic regression classifier.

1 Introduction

Recent work on word representation has been focusing on embedding syntactic and semantic information into fixed-sized vectors (Mikolov et al., 2013; Pennington et al., 2014) based on the distributional hypothesis, and have proven to be useful in many natural language tasks (Collobert et al., 2011). However, despite the rising popularity regarding the use of word embeddings, they often fail to capture the emotional semantics the words convey. For example, the GloVe vector captures the semantic meaning of “headache”, as it is closer to words of ill symptoms like “fever” and “toothache”, but misses the emotional association that the word carries. The word “headache” in the sentence “You are giving me a headache” does not really mean that the speaker will get a headache, but instead implies the negative emotion of the speaker.

To include affective information into the word representation, Tang et al. (2016) proposed Sentiment-Specific Word Embeddings (SSWE) which encodes both positive/negative sentiment and syntactic contextual information in a vector space. This work demonstrates the effectiveness of incorporating sentiment labels in a word-level information for sentiment-related tasks compared to other word embeddings. However, they only focus on binary labels, which weakens their generalization ability on other affect tasks. Yu et al. (2017) instead uses emotion lexicons to tune the vector space, which gives them better results. Nevertheless, this method requires human-labeled lexicons and cannot scale to large amounts of data. Felbo et al. (2017) achieves good results on affect tasks by training a two-layer bidirectional Long Short-Term Memory (bi-LSTM) model, named DeepMoji, to predict emoji of the input document using a huge dataset of 1.2 billions of tweets. However, collecting billions of tweets is expensive and time consuming for researchers.

Furthermore, most works in sentiment and emotion analysis have focused solely on a single task. Nevertheless, as emotion is a complex concept, we believe that all emotion involving situations such as stress, hate speech, sarcasm, and insult, should be included for a deeper understanding of emotion. Thus, one way to achieve this is through a multi-task training framework, as we present here.

Contributions: 1) We propose Emo2Vec [1] which are word-level representations that encode emotional semantics into fixed-sized, real-valued vectors. 2) We propose to learn Emo2Vec with a multi-task learning framework by including six different emotion-related tasks. 3) Compared to existing affect-related embeddings, Emo2Vec achieves better results on more than ten datasets with much less training data (1.9M vs 1.2B documents). Furthermore, with a simple logistic regression classifier, Emo2Vec reaches competitive performance to state-of-the-art results on several

[1] https://github.com/pxuab/emo2vec_wassa_paper

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 292–298
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

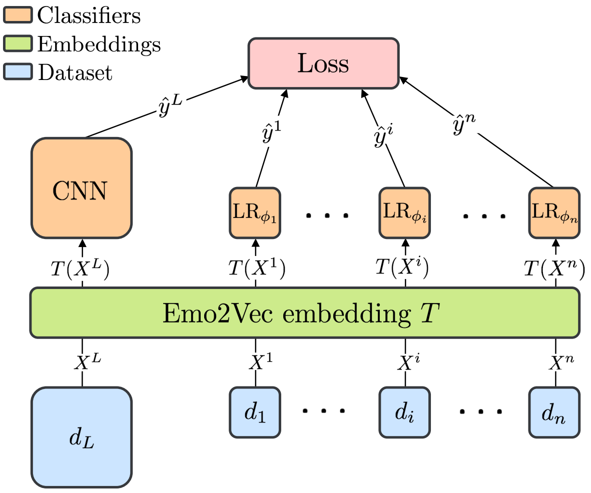

Figure 1: Multi-task learning diagram

datasets when combined with GloVe.

2 Methodology

We train Emo2Vec using an end-to-end multi-task learning framework with one larger dataset and several small task-specific datasets. The model is divided into two parts: a shared embedding layer (i.e. Emo2Vec), and task-specific classifiers. All datasets share the same word-level representations (i.e. Emo2Vec), thus forcing the model to encode shared knowledge into a single matrix. For the larger dataset, a Convolutional Neural Network (CNN) (LeCun et al., 1998) model is used to capture complex linguistic features present in the corpus. On the other hand, the classifier of each small dataset is a simple logistic regression.

Notation: We define $D = \{d_L, d_1, d_2, \cdots, d_n\}$ as the set of $n+1$ datasets, where d_L is the larger dataset and the other d_i are the small datasets. We denote a sentence X^i with $i \in \{L, 1, 2, \cdots, n\}$ as $[w_{i,1}, w_{i,2}, \cdots, w_{i,N_i}]$ where $w_{i,j}$ is the j-th word in the i-th sample and N_i is the number of words. All the models' parameters are defined as $M_\Phi = \{T, \text{CNN}, \text{LR}_{\phi_1}, \ldots, \text{LR}_{\phi_n}\}$, where $T \in \mathbb{R}^{|V| \times k}$ is the Emo2Vec matrix, $|V|$ is the vocabulary size and k is the embedding dimension, CNN is a Convolutional Neural Network model and LR_{ϕ_i} for $i \in [1, n]$ is a logistic regression classifier parameterized by ϕ_i which is specific for the dataset d_i. We denote the embedded representation of a word $w_{i,j}$ with $e_{w_{i,j}}$.

2.1 CNN model

The CNN architecture used is illustrated in Figure 2. Firstly, 1-D convolution is used to extract n-gram features from the input embeddings. Specifically, the j-th filter denoted by F_j, is convolved with embeddings of words in a sliding window of size k_j, giving a feature value $c_{j,t}$. J filters are learned trough this process:

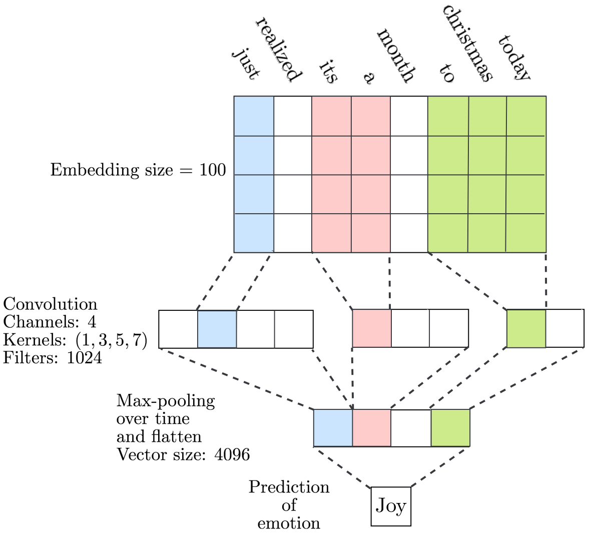

Figure 2: Structure of CNN model

$$c_{j,t} = F_j * e_{w_{L,t:t+k_j-1}} + b_j$$

where $*$ is the 1-D convolution operation. This is followed by a layer of ReLU activation (Nair and Hinton, 2010) for non-linearity. After that, we add a max-pooling layer of pooling size $M - F_j + 1$ along the time dimension to force the network to find the most relevant feature for predicting y^L correctly. The result of this series of operations is a scalar output of fm_j. All fm_j for $j \in [1, J]$ are then concatenated together to produce a vector representation $fm_{1:J}$ of the whole input sentence.

$$fm_j = \text{Max\_Pooling}\left(\text{ReLU}\left(c_{j,t}\right)\right)$$

To make the final classification, the vector $fm_{i,1:J}$ is projected to the target label space by adding another fully connected layer (i.e. parameterized by W and b), with a softmax activation.

$$\hat{y}^L = \text{Softmax}(W \cdot [fm_{1:J}] + b)$$

2.2 Multi-task learning

Since collecting a huge amount of labeled datasets is expensive, we collect two types of corpora, one larger dataset (millions of training samples) and a set of small datasets (thousands of training samples each) with accurate labels. For small datasets, sentiment analysis, emotion classification, sarcasm detection, abusive language classification, stress detection, insult classification and personality recognition are included. The reason why we include many datasets is to 1) leverage different aspects of words emotion knowledge, which may not be present in single domain dataset; 2) create a more general embedding emotional space that can generalize well across different tasks and domains. To avoid over-fitting, L2 regularization penalty is added from the weights of all logistic regression classifiers ϕ_i for $i \in [1, n]$. Hence, we jointly optimize the following loss function:

$$L(M_\Phi) = \frac{1}{n}\sum_{j=1}^{n} L_j + \lambda \sum_{j=1}^{n} \|\text{LR}_{\phi_j}\|_2$$

Where L_j is the negative log likelihood (NLL) between $\hat{y}^j$ and y^j, and λ an hyper-parameter for the regularization terms.

3 Experimental Setup

3.1 Dataset

Larger dataset

We collect a larger dataset from Twitter with hashtags as distant supervision. Such distant supervision method using hashtags has already been proved to provide reasonably relevant emotion labels by previous works (Wang et al., 2012).We construct our hashtag corpus from Wang et al. (2012), and Sintsova et al. (2017) [2]. More tweets between January and October 2017 are additionally added using the Twitter Firehose API by using the hashtags based on the hierarchy mentioned in Shaver et al. (1987). The hashtags are transformed into corresponding emotion labels of Joy, Sadness, Anger, and Fear. When extending the dataset, we only use documents with emotional hashtags at the end and filter out any documents with URLs, quotations, or less than five words as Wang et al. (2012) did. The total number of documents is about 1.9 million with four classes: joy (36.5%), sadness (33.8%), anger (23.5%), and fear (6%). The dataset is randomly split into a train (70%), validation (15%), and test set (15%) for experiments.

Small datasets

For sentiment, we include 8 datasets. (1,2) SST-fine and SST-binary (Socher et al., 2013) (3) OpeNER (Agerri et al., 2013) (4,5) tube_auto and tube_tablet (Uryupina et al., 2014) (6) SemEval (Hltcoe, 2013) (7,8) SS-Twitter and SS-Youtube (Thelwall et al., 2010). For emotion tasks, we include 4 datasets, (1) ISEAR (Wallbott and Scherer, 1986) (2) WASSA (Mohammad and Bravo-Marquez, 2017) (3) Olympic Sintsova et al. (2013) (4) SE0714 (Staiano and Guerini, 2014). We further include 6 other affect-related datasets. (1,2) SCv1-GEN and SCv2-GEN for sarcasm detection, (3) Stress (Winata et al., 2018), (4) Abusive (Waseem, 2016; Waseem and Hovy, 2016). (5) Personality (Pennebaker and King, 1999) (6) Insult. The detailed statistics can be found in Table 4 and Table 5 in Supplemental Material.

3.2 Pre-training Emo2Vec

Emo2Vec embedding matrix and the CNN model are pre-trained using hashtag corpus alone. Parameters of T and CNN are randomly initialized and Adam is used for optimization. Best parameter settings are tuned on the validation set. For the best model, we use the batch size of 16, embedding size of 100, 1024 filters and filter sizes are 1,3,5 and 7 respectively. We keep the trained embedding and rename it as CNN embedding for comparison. 100-dim for Emo2Vec is used in all experiments.

3.3 Multi-task training

We tune our parameters of learning rate, L2 regularization, whether to pre-train our model and batch size with the average accuracy of the development set of all datasets. We early stop our model when the averaged dev accuracy stop increasing. Our best model uses learning rate of 0.001, L2 regularization of 1.0, batch size of 32. We save the best model and take the embedding layer as Emo2Vec vectors.

3.4 Evaluation

Baselines: We use 50-dimension Sentiment-specific Word Embedding (SSWE) (Tang et al., 2016) as our baseline, which is an embedding model trained with 10 millions of tweets by encoding both semantic and sentiment information into vectors. Also, lots of work about the detection/classification in sentiment analysis implicitly encodes emotion inside the word vectors. For example, Felbo et al. (2017) trains a two-layer bidirectional Long Short-Term Memory (bi-LSTM) model, named DeepMoji, to predict emoji of the

[2] http://hci.epfl.ch/sharing-emotion-lexicons-and-data#emo-hash-data

model	SS-T	SS-Y	SS-binary	SS-fine	OpeNER	tube_auto	tube_tablet	SemEval	average
SSWE	**0.815**	0.835	0.698	0.365	0.701	0.620	0.654	0.629	0.665
DeepMoji embedding	0.788	0.841	0.751	0.369	**0.754**	0.628	0.675	**0.676**	0.685
CNN embedding	0.803	**0.862**	0.734	0.369	0.713	0.605	0.667	0.622	0.672
Emo2Vec	0.801	0.859	**0.812**	**0.416**	0.744	**0.629**	**0.688**	0.638	**0.698**

Table 1: Comparison between different emotion representations on sentiment datasets, all results are reported with accuracy. The best results are highlighted with bold fonts. Emo2Vec achieves best average score.

model	ISEAR	WASSA	SE0714	Olympic	Stress	SCv1-GEN	SCv2-GEN	Insult	Abusive	Personality	average
SSWE	0.327	0.466	0.217	0.508	0.704	0.660	0.678	0.559	0.539	0.674	0.533
DeepMoji embedding	0.379	0.532	0.286	0.485	0.739	0.658	0.685	**0.666**	0.586	**0.678**	0.569
CNN embedding	**0.384**	0.549	0.259	0.480	**0.744**	0.657	0.707	0.623	0.560	0.676	0.564
Emo2Vec	0.372	**0.559**	**0.323**	**0.506**	**0.744**	**0.674**	**0.710**	0.647	**0.588**	0.675	**0.580**

Table 2: Comparison between different representations on other affect related datasets. All results are reported with f1 score. The best results are highlighted with bold fonts. On average, Emo2Vec achieves best f1 score.

input document using a huge dataset of 1.2 billion tweets. Their embedding layer is implicitly encoded with emotion knowledge. Thus, we use the DeepMoji embedding, the 256-dimension embedding layer of DeepMoji as another baseline.

Evaluation method: To make a fair comparison with other baseline representations, we first take one dataset d_i out from n small datasets as the test set. The remaining $n - 1$ small datasets and the larger dataset are used to train Emo2Vec through multi-task learning. We take the trained Emo2Vec as the feature for d_i and train a logistic regression on d_i to compare the performance with other baseline representations. The procedure is repeated n times to see the generalization ability on different datasets. We release Emo2Vec trained on all datasets. For sentiment tasks, accuracy score is reported. For other tasks, if it is binary task, we report f1 score for the positive class. If it is multi-class classification tasks, we make it binary classification problem for each class and report averaged f1 score.

4 Results

We compare our Emo2Vec with SSWE, CNN embedding, DeepMoji embedding and state-of-the-art(SOTA) results on 18 different datasets. The results can be found in Table 1 and Table 2.

Compared with CNN embedding: Emo2Vec works better than CNN embedding on 14/18 datasets, giving 2.6% absolute accuracy improvement for the sentiment task and 1.6% absolute f1-score improvement on the other tasks. It shows multi-task training helps to create better generalized word emotion representations than just using a single task.

Compared with SSWE: Emo2Vec works much better on all datasets except SS-T datasets, which gives 3.3% accuracy improvement and 4.7% f1 score improvement respectively on sentiment and other tasks. This is because SSWE is trained on 10M binary classification task on twitter which then over-fits on dataset SS-T, and generalizes poorly to other tasks.

Compared with DeepMoji embedding: Emo2Vec outperforms DeepMoji on 13/18 datasets despite the much smaller size of our training corpus (1.9M documents for us vs 1.2B documents for DeepMoji). On average, it gives 1.3% improvement in accuracy for the sentiment task and 1.1% improvement of f1-score on the other tasks.

Compared with SOTA results: We further compare the performance of Emo2Vec vectors with SOTA results on 14 datasets where the same split is shared. Since Emo2Vec is not trained by predicting contextual words, it is weak on capturing synthetic and semantic meaning. Thus, we concatenate Emo2Vec with the pre-trained GloVe vectors, which are trained on Twitter and Wikepedia [3]. Then, the concatenated vector of GloVe and Emo2Vec, the concatenated vector of GloVe and DeepMoji embeddings and GloVe are included for comparison with SOTA results. Note that SOTA results require complex bi-LSTM model while all these representations are trained and reported with a logistic regression classifier. Here, we want to highlight that solely using a simple classifier with good word representation can achieve promising results.

Table 3 shows that GloVe+Emo2Vec outperforms GloVe on 13/14 datasets. Compared with GloVe+DeepMoji, GloVe+Emo2Vec achieves same or better results on 11/14 datasets, which on average gives 1.0% improvement. GloVe+Emo2Vec achieves better performances on

[3] http://nlp.stanford.edu/data/glove.twitter.27B.zip and http://nlp.stanford.edu/data/glove.6B.zip

dataset	Previous SOTA results		GloVe	GloVe+DeepMoji	GloVe+Emo2Vec
SS-Twitter	bi-LSTM (Felbo et al., 2017)	0.88	0.78	**0.81**	**0.81**
SS-Youtube	bi-LSTM (Felbo et al., 2017)	0.93	0.84	0.86	**0.87**
SS-binary	bi-LSTM (Yu et al., 2017)	0.886	0.795	0.809	**0.823**
SS-fine	bi-LSTM (Yu et al., 2017)	0.497	0.414	0.421	**0.436**
OpeNER	bi-LSTM (Barnes et al., 2017)	0.825	0.750	**0.781**	0.778
tube_auto	SVM (Barnes et al., 2017)	0.662	0.630	0.628	**0.660**
tube_tablet	SVM (Barnes et al., 2017)	0.681	0.650	0.678	$\boxed{\mathbf{0.684}}$
SemEval	bi-LSTM (Barnes et al., 2017)	0.685	0.671	**0.695**	0.680
ISEAR	bi-LSTM (Felbo et al., 2017)	0.57	0.41	0.43	**0.45**
SE0714	bi-LSTM (Felbo et al., 2017)	0.37	0.36	0.36	$\boxed{\mathbf{0.43}}$
Olympic	bi-LSTM (Felbo et al., 2017)	0.61	0.52	0.52	**0.53**
stress	bi-LSTM (Winata et al., 2018)	0.743	0.759	**0.793**	0.770
SCv1-GEN	bi-LSTM (Felbo et al., 2017)	0.69	**0.69**	0.68	0.68
SCv2-GEN	bi-LSTM (Felbo et al., 2017)	0.75	0.73	**0.74**	**0.74**
Average			0.642	0.657	**0.667**

Table 3: Comparison between different word-level emotion representations with state-of-the-art results. The best results are in bold. New state-of-the-art results Emo2Vec that achieves are highlighted with boxes.

SOTA results on three datasets (SE0714, stress and tube_tablet) and comparable result to SOTA on another four datasets (tube_auto, SemEval, SCv1-GEN and SCv2-GEN). We believe the reason why we achieve a much better performance than SOTA on the SE0714 is that headlines are usually short and emotional words exist more commonly in headlines. Thus, to detect the corresponding emotion, more attention needs to be paid to words.

5 Related work

For sentiment analysis, numerous classification models (Kalchbrenner et al.; Iyyer et al., 2015; Dou, 2017) have been explored. Multi-modal sentiment analysis (Zadeh et al., 2017; Poria et al., 2017) extends text-based model to the combination of visual, acoustic and language, which achieves better results than the single modality. Various methods are developed for automatic constructions of sentiment lexicons using both supervised and unsupervised way (Wang and Xia, 2017). Aspect-based sentiment (Chen et al., 2017; Wang et al., 2016) is also a hot topic where researchers care more about the sentiment towards a certain target. Transfer learning from the large corpus is also investigated by Felbo et al. (2017) to train a large model on a huge emoji tweet corpus, which boosts the performance of affect-related tasks. Multi-task training has achieved great success in various natural language tasks, such as machine translation (Dong et al., 2015; Malaviya et al., 2017), multilingual tasks (Duong et al., 2015; Gillick et al., 2016), semantic parsing (Peng et al., 2017). Hashimoto et al. (2017) jointly learns POS tagging, chunking, dependency parsing, semantic relatedness, and textual entailment by considering linguistic hierarchy and achieves state-of-the-results on five datasets. For sentiment analysis, Balikas et al. (2017) jointly trains ternary and fine-grained classification with a recurrent neural network and achieves new state-of-the-art results.

6 Conclusion and Future Work

In this paper, we propose Emo2Vec to represent emotion with vectors using a multi-task training framework. Six affect-related tasks are utilized, including emotion/sentiment analysis, sarcasm classification, stress detection, abusive language classification, insult detection, and personality recognition. We empirically show how Emo2Vec leverages multi-task training to learn a generalized emotion representation. In addition, Emo2Vec outperforms existing affect-related embeddings on more than ten different datasets. By combining Emo2Vec with GloVe, logistic regression can achieve competitive performances on several state-of-the-art results.

7 Acknowledgements

This work is partially funded by ITS/319/16FP of Innovation Technology Commission, HKUST 16248016 of Hong Kong Research Grants Council.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51).

Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask learning for fine-grained twitter sentiment analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1005–1008, New York, NY, USA. ACM.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461.

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

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Zi-Yi Dou. 2017. Capturing user and product information for document level sentiment analysis with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 521–526.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 845–850.

Bjarke Felbo, Alan Mislove, Anders Sgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1616–1626.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306.

Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. 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.

J Hltcoe. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. *Atlanta, Georgia, USA*, 312.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Hao Peng, Sam Thomson, and Noah A Smith. 2017. Deep multitask learning for semantic dependency parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2037–2048.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Valentina Sintsova, Margarita Bolvar Jimnez, and Pearl Pu. 2017. Modeling the impact of modifiers on emotional statements. In *Proceedings of the 18th Int. Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Valentina Sintsova, Claudiu-Cristian Musat, and Pearl Pu. 2013. Fine-grained emotion recognition in olympic tweets based on human computation. In *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 427–433.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558.

Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249. Citeseer.

Harald G. Wallbott and Klaus R. Scherer. 1986. How universal and specific is emotional experience? evidence from 27 countries on five continents. *Information (International Social Science Council)*, 25(4):763–795.

Leyi Wang and Rui Xia. 2017. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 502–510.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

A Supplemental Material

A.1 Preprocessing

Numbers and user mentions are changed to special tokens (e.g. <number>, <user>), and other non-alphanumeric characters are removed to reduce noisy characters from the corpus. All words are lowercased. The tweet tokenizer from NLTK (Bird and Loper, 2004) library used to tokenize each document. We use the vocabulary of GloVe twitter for its huge size of 1193514 tokens. We further add "UNK" to denote out-of-vocabulary words and "PAD" to pad sentences.

datasets	SS-T	SS-Y	SS-binary	SS-fine	OpeNER	tube_auto	tube_tablet	SemEval
Domain	Tweets	Coments	Movie reviews	Movie reviews	Hotel reviews	Coments	Coments	Tweets
#classes	2	2	2	5	4	2	2	3
#training	800	800	6920	8544	2780	3381	4997	6021
#validation	100	100	872	1102	186	225	333	890
#testing	1213	1242	1821	2210	743	903	1334	2376

Table 4: statistics of 8 sentiment datasets

datasets	ISEAR	WASSA	SE0714	Olympic	Stress	SCv1-GEN	SCv2-GEN	Insult	Abusive	Personality
Domain	Experiences	tweets	Headlines	Tweets	Interviews	Debates	Debates	Comments	Tweets	Essays
#classes	7	4	3	4	2	2	2	2	2	5
#training	900	3613	200	200	1751	800	800	4450	12016	1578
#validation	100	347	50	50	200	100	100	495	3005	395
#testing	6480	3142	1000	762	320	1095	2360	1237	3756	494

Table 5: statistics of emotion and other datasets