eMLM: A New Pre-training Objective for Emotion Related Tasks

Tiberiu Sosea
Computer Science
University of Illinois at Chicago
tsosea2@uic.edu

Cornelia Caragea
Computer Science
University of Illinois at Chicago
cornelia@uic.edu

Abstract

Bidirectional Encoder Representations from Transformers (BERT) have been shown to be extremely effective on a wide variety of natural language processing tasks, including sentiment analysis and emotion detection. However, the proposed pre-training objectives of BERT do not induce any sentiment or emotion-specific biases into the model. In this paper, we present Emotion Masked Language Modeling, a variation of Masked Language Modeling, aimed at improving the BERT language representation model for emotion detection and sentiment analysis tasks. Using the same pre-training corpora as the original BERT model, Wikipedia and BookCorpus, our BERT variation manages to improve the downstream performance on 4 tasks for emotion detection and sentiment analysis by an average of 1.2% F1. Moreover, our approach shows an increased performance in our task-specific robustness tests. We make our code and pre-trained model available at https://github.com/tsosea2/eMLM.

1 Introduction

Language models have been studied extensively in the NLP community (Dai and Le, 2015; Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019), with approaches attaining state-of-the-art results on multiple token-level or sentence-level tasks. BERT (Devlin et al., 2019) is a pre-trained language model, which proposed a new pre-training objective inspired by the Cloze task (Taylor, 1953), which enables the training of a deep bi-directional transformer network. This objective, called Masked Language Modeling (MLM) is used on large amounts of unlabeled data from Wikipedia and BookCorpus to produce powerful universal language representations. However, the pre-training does not take into account the downstream task on which the model will be applied.

In this paper, we posit that we can leverage the characteristics of a downstream task to design better task-tailored pre-training objectives. Concretely, we induce information from emotion or sentiment lexicons into our BERT pre-training objective to improve the performance on tasks from sentiment analysis and emotion detection.

There are numerous studies that focus on emotion detection (Demszky et al., 2020; Desai et al., 2020; del Arco et al., 2020; Sosea and Caragea, 2020; Majumder et al., 2019; Mohammad and Kiritchenko, 2018; Abdul-Mageed and Ungar, 2017; Mohammad and Kiritchenko, 2015; Mohammad, 2012; Strapparava and Mihalcea, 2008) and sentiment analysis (Yin et al., 2020; Tian et al., 2020; Phan and Ogunbona, 2020; Zhai and Zhang, 2016; Chen et al., 2016; Liu, 2012; Glorot et al., 2011; Pang and Lee, 2005). Various lexicons have been used to improve model performance on these tasks. For instance, Katz et al. (2007) used occurrences of emotion words to identify various emotion types in news headlines. Moreover, emotion lexicons have been used to produce important features which can be used inside a machine learning algorithm to improve the performance on emotion detection tasks (Mohammad, 2012; Sykora et al., 2013; Khanpour and Caragea, 2018; Biyani et al., 2014). In this paper, however, instead of leveraging these lexicons to design features, in contrast, we use them to obtain language representations that are more suitable for emotion and sentiment tasks.

To this end, we introduce Emotion Masked Language Modeling (eMLM), a new pre-training BERT (Devlin et al., 2019) objective aimed at improving the BERT performance on tasks related to sentiment analysis and emotion detection. Inspired by the well-known Masked Language Modeling objective, eMLM adds only a few simple, yet powerful changes. Instead of uniformly masking the tokens in the input sequence, eMLM leverages

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 286–293
August 1–6, 2021.

SENT	They	look	absolutely	**perfect**	together	I	**hope**	its	that	way	in	real	life	too
MLM	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
eMLM	0.09	0.09	0.09	**0.50**	0.09	0.09	**0.50**	0.09	0.09	0.09	0.09	0.09	0.09	0.09
SENT	Most	**tiring**	thing	was	the	drive	one	hour	each	way				
MLM	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15				
eMLM	0.11	**0.50**	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11				

Table 1: Comparison of masked probabilities between MLM and eMLM on two example sentences.

lexicon information, and assigns higher masking probabilities to words that are more likely to be important in the sentiment or emotion contexts. To enable a fair comparison with the vanilla BERT model, we train the eMLM BERT model in the same fashion as the vanilla BERT, pre-training on Wikipedia and BookCorpus (Zhu et al., 2015). To our knowledge, we are the first to study different masking probabilities for the BERT pre-training procedure guided by sentiment and emotion lexicons. Similar to our work, some studies also focused on incorporating sentiment information into pre-trained language models. For example, Yin et al. (2020) built an attention network on top of BERT to predict sentiment labels of phrase nodes obtained through a constituency parse tree. On the other hand, Tian et al. (2020) designed various pre-training objectives, such as masking and predicting all words from a pre-defined small set of seeds, and predicting an aspect-sentiment pair or the polarity of words. In contrast, we leverage information from available sentiment and emotion lexicons.

We show the feasibility of our approach by testing eMLM on two sentiment analysis benchmark datasets and two emotion detection datasets. These datasets span diverse domains, such as movie reviews, online health communities, and Reddit discussions, enabling a comprehensive analysis of eMLM.

Our contributions are as follows: **1)** We introduce a new pre-training objective for BERT (leveraging available lexicons), aimed at producing better task-guided universal representations for downstream tasks from sentiment analysis and emotion detection. We offer the pre-trained model as an easy way to leverage our approach on downstream applications. **2)** We show the efficacy of our approach by testing our method on four benchmark datasets for emotion and sentiment and obtain an average improvement in F1 score of 1.2%. **3)** We verify the robustness of our model in the face of input perturbations, which occur frequently in informal contexts (e.g., due to mispellings).

2 Proposed Approach

Background Bidirectional Encoder from Transformers for Language Understanding (BERT) (Devlin et al., 2019) is a pre-trained language model trained on large amounts of unlabeled data using two objectives: **1)** Masked Language Modeling (MLM) randomly masks 15% of tokens in a sequence, followed by a supervised prediction of the masked tokens; **2)** Next Sentence Prediction (NSP) predicts in a binary fashion if two sentences follow each other. By using these two tasks on large-scale data repositories such as BookCorpus (800M words) (Zhu et al., 2015) and Wikipedia ($2,500$M words), BERT produces powerful universal language representations, applicable on a wide range of tasks, such as sentiment analysis, question answering, and commonsense reasoning.

However, to be used in various downstream tasks, BERT has to undergo a task-specific fine-tuning step (Devlin et al., 2019), where the contextualized embedding is adapted to the needed task. We posit that we can improve the downstream performance by focusing on the target task in the pre-training phase as well. Specifically, we focus on sentiment analysis and emotion detection, and show that task-guided unsupervised pre-training helps the performance considerably.

Masking Emotion Words Now we introduce Emotion Masked Language Modeling (eMLM), a variation of MLM targeted at inducing emotion or sentiment-specific biases in the BERT pre-training phase. Specifically, unlike BERT, which uses a uniform probability (15%) to mask the tokens in an input sentence, we assign higher probabilities to tokens which are emotionally rich words from an available lexicon $\mathcal{L}$. We denote this probability by k, which is a hyperparameter in our eMLM method. Our masking process can be summarized as follows: Given an input sentence S: **1)** We extract the words that belong to the lexicon $\mathcal{L}$, and we denote them by E; **2)** We set the masking probability of these words as $P(w_e) = k \; \forall \; w_e \in E$; **3)** To ensure

we mask 15% of the words in total, we lower the masking probability of the non-emotionally-rich words using the following formula:

$$P(w_n) = \frac{max(|S| \cdot 0.15 - |E| \cdot k,\ 0)}{|S| - |E|},\ \forall\ w_n \notin E$$

where $|\cdot|$ represents the size of a set. We show examples of how our masked probabilities change from MLM to eMLM in Table 1. For instance, in the first example, there are two emotion words, *perfect* and *hope*, and we use a masking probability of $k = 0.50$. While the probabilities of these two words are set to 50%, the non emotionally-rich word probability is lowered from 15% to 9% to keep the sum of probabilities constant. The rest of the training process is the same as the original BERT pre-training. That is, we train our BERT model from scratch using eMLM and NSP on the same datasets: Wikipedia and BookCorpus. We mention that we use whole word masking, both for eMLM and the MLM (i.e., we mask all the subtokens corresponding to a word).

3 Experiments and Results

In this section, we first describe our experimental setup (§3.1), then present our datasets and lexicons (§3.2), and then discuss the results that contrast **eMLM** with the original BERT **MLM** (§3.3).

3.1 Experimental Setup

We use various benchmark datasets from sentiment analysis and emotion detection to test our eMLM approach. For every dataset considered, we use the provided training, validation, and test splits. To assert statistical significance, we fine-tune each model 10 times with different random seeds and report the average F1 score. We investigate various masking probabilities k, ranging from 0.2 to 1.0, and find that 0.5 works best in our setting. For low values around 0.2 we notice that the performance is similar to that of the original BERT, while for high values (closer to 1.0), the performance is negatively affected.

3.2 Datasets and Lexicons

We test our models on various benchmark datasets described below.

Stanford Sentiment Treebank (**SST**) (Socher et al., 2013) SST contains 11,855 sentences from movie reviews, annotated with five sentiment labels: *negative*, *somewhat negative*, *neutral*, *somewhat positive*, and *positive*. First, we consider the binarized dataset, called SST-2, where the examples with the *negative* and *somewhat negative* labels are merged into a *negative* class, and the examples with the *somewhat positive* and *positive* labels are merged into a *positive* class (with neutral class being removed). Second, we consider the SST fine-grained version (SST-5), which uses all five labels.

	SST-2		SST-5	
	ACC	F-1	ACC	F-1
BERT	0.912	0.922	0.532	0.541
eMLM (S)	0.919	0.928	0.541	0.552
eMLM (E)	**0.920**	**0.931**†	**0.547**	**0.558**†

Table 2: Performance on the sentiment analysis task. We assert significance† if $p < 0.05$ under a t-test with the vanilla BERT model.

GoEmotions (Demszky et al., 2020) is a sentence-level multilabel dataset of 58,000 comments curated from Reddit and annotated with 27 emotion categories and the neutral class.

CancerEmo (Sosea and Caragea, 2020) is a sentence-level multilabel dataset of 8,500 sentences labeled with the eight Plutchik (Plutchik, 1980) basic emotions from an Online Health Community for people suffering from diseases such as cancer.

We analyze the behaviour of eMLM in diverse environments: sentiment analysis or emotion detection, various data platforms (e.g., Reddit, OHCs), and variate emotion or sentiment granularity (from 2 classes to as many as 28 classes).

Lexicons As mentioned above, our eMLM focuses on emotionally rich words from a lexicon. In this paper, we use EmoLex (Mohammad and Turney, 2013), a lexicon of 6,000 words associated with eight Plutchik basic emotions (Plutchik, 1980) (sadness, anger, joy, surprise, anticipation, trust, fear, disgust) and 5,555 words associated with the positive and negative sentiments. We consider the sentiment and emotion words separately to analyze the impact of each on the performance of eMLM. We denote the approach which masks the emotion-revealing words by eMLM (E), and the sentiment-revealing words by eMLM (S).

EMOTION	BERT	eMLM (E)	eMLM (S)
ADMIRATION	0.65	**0.68**†	0.67
AMUSEMENT	0.80	**0.83**†	0.82
ANGER	**0.47**	0.46	0.46
ANNOYANCE	0.34	0.34	0.34
APPROVAL	0.36	**0.38**	0.37
CARING	0.39	**0.43**	0.42
CONFUSION	0.37	0.37	0.37
CURIOSITY	0.54	**0.57**†	0.57
DESIRE	0.49	0.49	0.49
DISAPPOINTMENT	0.28	**0.30**	**0.30**
DISAPPROVAL	0.39	**0.43**†	0.41
DISGUST	0.45	**0.48**†	0.48
EMBARRASSMENT	0.43	0.43	**0.44**
EXCITEMENT	0.34	0.34	0.34
FEAR	0.60	**0.64**†	0.63
GRATITUDE	0.86	**0.88**†	0.87
GRIEF	0.00	0.00	0.00
JOY	0.51	**0.53**	0.52
LOVE	0.78	**0.80**†	**0.80**
NERVOUSNESS	0.35	**0.37**	0.36
NEUTRAL	**0.68**	0.67	**0.68**
OPTIMISM	0.51	**0.53**	0.52
PRIDE	0.36	0.36	0.36
REALIZATION	0.21	0.21	0.21
RELIEF	0.15	**0.16**	**0.16**
REMORSE	**0.66**	0.65	**0.66**
SADNESS	**0.49**	**0.49**	0.48
SURPRISE	0.50	**0.53**†	0.52
AVERAGE	0.462	**0.476**	0.469

Table 3: F-1 scores on the Goemotion dataset. We assert significance† if $p < 0.05$ under a t-test with the vanilla BERT model.

3.3 Results

Results on Sentiment Analysis We show the results of our approaches on SST in Table 2. First, we observe that eMLM (E) and eMLM (S) improve upon the vanilla BERT model on both tasks, with eMLM (E) obtaining as much as 1.7% improvement in F1. Interestingly, eMLM (E) outperforms eMLM (S) suggesting that masking finer-granularity emotion words in eMLM produces better representations for the task. At the same time, eMLM (E) achieves better performance on the fine-grained SST-5 task, where the improvements over the vanilla BERT are considerable.

Results on Emotion Detection We show the results of eMLM on the GoEmotions dataset in Table 3 and observe that, similar to sentiment analysis, eMLM (E) is the best performing approach, improving upon vanilla BERT by 1.4% in F1. We show the results on CancerEmo in Table 4 and observe the same pattern: **eMLM (E) consistently outperforms the other approaches**. We see improvements as high as 4% on Joy and 2% on Sadness. Overall, eMLM (E) obtains an 1.7% F1 improvement over the vanilla BERT model.

EMOTION	BERT	eMLM (E)	eMLM (S)
SADNESS	0.71	**0.73**†	**0.73**†
JOY	0.81	**0.85**†	0.84
FEAR	**0.77**	**0.77**	**0.77**
ANGER	0.68	**0.69**	**0.69**
SURPRISE	**0.68**	**0.68**	0.67
DISGUST	**0.59**	0.58	0.57
TRUST	0.67	0.67	0.67
ANTICIPATION	0.70	**0.78**†	0.74
AVERAGE	0.701	**0.718**	0.706

Table 4: Performance on CancerEmo dataset. We assert significance† if $p < 0.05$ under a t-test with the vanilla BERT model.

K	SST-2	SST-5	CANCEREMO	GOEMOTIONS
0.15	0.922	0.541	0.701	0.462
0.30	0.923	0.540	0.704	0.466
0.50	0.931	0.558	0.718	0.476
0.70	0.921	0.539	0.700	0.455
0.90	0.911	0.540	0.691	0.412

Table 5: Average F-1 on the considered datasets using various values of the emotion masking probability k.

Discussion The presented results reveal the feasibility of our proposed approach. Our BERT model trained using the eMLM objective produces high quality contextualized embeddings for downstream tasks that span the sentiment analysis and emotion detection tasks. Moreover, our methods incur no additional computational cost over the original BERT (Devlin et al., 2019), and undergo the same amount of pre-training. We also tried combining and masking both sentiment and emotion words; however, we did not see any performance improvements. As a step forward, we are interested in gaining more insights into the differences between eMLM (E) and the vanilla BERT model. We study this in the robustness context in the next section, and analyze how our models behave in the face of various input perturbations (i.e., noise).

Varying the Emotion Masking Probability k To offer additional insights into our eMLM approach and show the impact of the sentiment or emotion-rich word masking probability on downstream tasks, we show the results obtained using various values of k in Table 5. First, we note that using a slightly lower probability of 0.30 still adds improvements to our model on three of the considered datasets. In contrast, too high of a proba-

bility hurts the F1 performance. Concretely, using $k = 0.90$, our eMLM approach decreases the F1 compared to the vanilla BERT by 1% on **CancerEmo**, 5% on GoEmotions, and 1% on **SST-2**.

4 Robustness Test

It has been shown that neural models are often sensitive to various input perturbations (Niu et al., 2020; Belinkov and Bisk, 2018). In this section, we aim to investigate the robustness of our proposed approach in the face of input noise. We focus on the following two questions: **1)** Does eMLM improve the robustness of the model? **2)** What type of input noise is successful in misleading our model? We study these questions on the SST-5 sentiment analysis task using the framework introduced by Hsieh et al. (2019). We explore three ways to generate input perturbations and verify their "success." We say a perturbation is "successful" on a model M for an example e if **1)** The model M classifies e correctly and **2)** The model M misclassifies the example e when noise is applied to it. Naturally, the lower the perturbation success rate, the more robust a model is. The perturbations that we considered are as follows:

1. **Random** (Alzantot et al., 2018) replaces one word from the input sentence with a random word from the vocabulary. For a word, we repeat this process 100 times. If at least one of the replacements leads to an incorrect prediction, the perturbation is deemed to be successful.

2. **LIST** (Alzantot et al., 2018) replaces each word (one at a time) in the input text with a synonym. The input perturbation is successful if at least one replacement leads to an incorrect prediction.

3. **EmoWord** If there is an emotion word in the input sentence, then we zero out that word, otherwise, we zero out a random word from the input sequence.

Results We show the results of the robustness tests for the vanilla BERT and the eMLM approach in Table 6. First, EmoWord is the most successful perturbation, being twice as effective compared to the other methods. Second, we observe that Random and LIST obtain the same success rates among both the BERT and eMLM approach. However, on EmoWord, our eMLM approach is considerably more robust, outperforming the simple BERT model by 4.4%. We argue that this is the byproduct of the eMLM training procedure, which focuses on predicting emotion words in the pre-training step.

EMOTION	RANDOM	LIST	EMOWORD
BERT	1.5%	2.4%	9.8%
eMLM	1.5%	2.4%	**5.4%**

Table 6: Robustness of our models in terms of perturbation success rates. Lower success rates indicate more robust models.

5 Conclusion

In this paper, we introduced a new BERT pre-training objective suited for sentiment analysis and emotion detection tasks. We showed that the approach is feasible; it needs no additional pre-training compared to the vanilla BERT, and improves the performance by 1.2% F1 on average on various tasks. Our analysis also suggests that eMLM is more robust in the face of input perturbations. As future work, we note that our approach is general enough, so we plan to leverage different lexicons outside the sentiment analysis and emotion detection domains to investigate if the model generalizes well on other domains (e.g., financial). We also plan to study if our method is effective for non-English languages. Finally, we note that there exist lexicons that assign to words not only their emotion, but also their emotion intensity (Mohammad, 2018). Therefore, we plan to investigate if associating the masking probability with the emotion intensity (i.e., assign a higher probability to a more intensive word) would further help improve the performance.

Acknowledgments

We thank our anonymous reviewers for their constructive comments and feedback. This work is partially supported by the NSF Grants IIS-1912887 and IIS-1903963. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. The computation for this project was performed on Amazon Web Services through a research grant.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, Carlo Strapparava, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1492–1498. European Language Resources Association.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 827–836, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, Austin, Texas. Association for Computational Linguistics.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3079–3087. Curran Associates, Inc.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

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.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *In Proceedings of the Twenty-eight International Conference on Machine Learning, ICML*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, Florence, Italy. Association for Computational Linguistics.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. SWAT-MP:the SemEval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313, Prague, Czech Republic. Association for Computational Linguistics.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI*

2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6818–6825. AAAI Press.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Online. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, March 16-20, 2008*, pages 1556–1560. ACM.

Martin D Sykora, Thomas Jackson, Ann O'Brien, and Suzanne Elayan. 2013. Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *IADIS Int. J. Comput. Sci. Inf. Syst*, 2013:19–26.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.

Shuangfei Zhai and Zhongfei (Mark) Zhang. 2016. Semisupervised autoencoder for sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 1394–1400. AAAI Press.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies:

Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.