

Guilt by Association: Emotion Intensities in Lexical Representations

Shahab Raji

Rutgers University
shahab.raji@rutgers.edu

Gerard de Melo

Hasso Plattner Institute/University of Potsdam
<http://gerard.demelo.org>

Abstract

What do linguistic models reveal about the emotions associated with words? In this study, we consider the task of estimating word-level emotion intensity scores for specific emotions, exploring unsupervised, supervised, and finally a self-supervised method of extracting emotional associations from pretrained vectors and models. Overall, we find that linguistic models carry substantial potential for inducing fine-grained emotion intensity scores, showing a far higher correlation with human ground truth ratings than state-of-the-art emotion lexicons based on labeled data.

1 Introduction

There has been substantial research on methods to label words with associated emotions. Crowdsourcing approaches have been used to compile databases to study the nexus between them (Mohammad and Turney, 2013; Mohammad, 2018; Shoeb and de Melo, 2020). Another strategy, adopted by the well-known DepecheMood (Staiano and Guerini, 2014) and DepecheMood++ (Araque et al., 2018) lexicons is to apply simple statistical methods to emotionally tagged data crawled from specific online sources.

In this paper, we consider the question: What do linguistic models reveal about the emotions associated with words? Word vectors (Mikolov et al., 2013; Pennington et al., 2014) have often been evaluated on standard word relatedness benchmarks. In this paper, we instead explore to what extent they encode emotional associations.

Earlier methods (Strapparava and Mihalcea, 2008; Mac Kim et al., 2010) used corpus statistics in tandem with dimensionality reduction techniques for emotion prediction. Shoeb et al. (2019) considered word embeddings trained on emojis for emotion analysis. Rothe et al. (2016) proposed predicting sentiment polarity ratings from word vectors, while other studies (Buechel and Hahn,

2018b; Buechel et al., 2018) predicted valence, arousal, and dominance using supervised deep neural networks. Khosla et al. (2018) proposed incorporating valence, arousal, and dominance ratings as additional signals into dense word embeddings. Buechel and Hahn (2018a) showed that emotion intensity scores can be predicted based on a lexicon providing valence–arousal–dominance ratings. The task of emotional association of words has been studied on other languages as well. Sidorov et al. (2012) presents a dataset of Spanish words labelled with Ekman’s six emotions, while others explored cross-lingual propagation from one language to another (Abdaoui et al., 2017) or to hundreds of other languages (Ramachandran and de Melo, 2020). We show that we can make use of linguistic models to obtain high correlations with emotion intensities without any need for manual ratings. While this work focuses on English, the methodology can be applied to data in any language.

Overview. We introduce the task and experimental setup in Section 2. Then, section 3 presents our techniques to address this task based on linguistic models. We provide detailed empirical evaluation results in Section 4. Finally, Section 5 concludes the paper by discussing the relevance of our experiments.

Contribution. Overall, our intriguing finding in this paper is that pretrained linguistic models allow us to predict much more accurate emotion associations than state-of-the-art emotion lexicons. We show that this holds even without any supervision, while different kinds of supervised setups yield even better results.

2 Emotion Intensities as Associations

In our study, we seek to predict the emotions associated with individual words. Such predictions can also be useful for sentence-level predictions,

as considered in Section 4.2, but that is not the primary focus of this paper.

While many emotion lexicons only provide binary emotion labels, it is clear that words may exhibit different degrees of association with an emotion. For instance, the word *party* suggests a greater degree of *joy* than the word *sea*, although the latter may as well be associated with *joy*.

Emotion inventory and ground truth. The NRC Emotion/Affect Intensity Lexicon (Mohammad, 2018) provides ground-truth human ratings of emotion intensity for English words. The latest release of the data covers the eight basic emotions proposed in Plutchik’s wheel of emotions (Plutchik, 1980), i.e., *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*. However, the techniques we consider in our paper apply broadly to any inventory of discrete emotions.

The ratings in the dataset are scaled to $[0, 1]$ such that a score of 1 signifies that a word “conveys the highest amount of” a particular emotion, while 0 corresponds to the lowest amount. These ratings were solicited via crowdsourcing and then standardized using best–worst scaling.

Benchmark data. In our work, instead of viewing the resource as a lexicon that provides emotion intensity scores, we propose to treat it as providing gold standard *associations between pairs of words*, one of the two words being an emotion name. Thus, we derive a semantic association benchmark similar to widely used word relatedness benchmarks such as the RG-65 (Rubenstein and Goodenough, 1965) and WS353 (Finkelstein et al., 2001) sets.

In order to more fairly evaluate resources that only provide unigrams, we disregard bigrams in the lexicon, resulting in a total of 9,706 word pairs. To facilitate an evaluation of unsupervised and supervised techniques, for each emotion, we split the corresponding data into training/validation/test portions with a ratio of 64%/16%/20%. This results in a total of 7,762 pairs in the training sets, and 1,944 pairs in the test sets. The remaining 1,552 instances serve as validation data. We adopt these data splits for all of the experiments in this paper and also make them available online for reproducibility.¹

3 Estimating Emotion Intensity

Emotion lexicons are typically created using crowdsourcing (Mohammad and Turney, 2013) or draw-

ing on emotion-labeled text (Staiano and Guerini, 2014; Araque et al., 2018). We investigate deriving emotion scores from word embedding models without any emotion-specific signals.

3.1 Unsupervised Prediction

While past work on emotion intensity prediction has considered this an entirely separate task, we here consider emotional intensity scoring as similar in nature to regular lexical associations between words. Given two words w_1, w_2 , the cosine similarity $\cos(\mathbf{v}_{w_1}, \mathbf{v}_{w_2})$ of their corresponding word vectors $\mathbf{v}_{w_1}, \mathbf{v}_{w_2}$ is expected to reflect their association, as most word vector spaces capture semantic relatedness (Hill et al., 2014). Hence, given a targeted vocabulary word w and an emotion e under consideration, we simply compute

$$\sigma_u(w, e) = \cos(\mathbf{v}_w, \mathbf{v}_{w_e}) = \frac{\mathbf{v}_w \cdot \mathbf{v}_{w_e}}{|\mathbf{v}_w| |\mathbf{v}_{w_e}|} \quad (1)$$

as the emotion intensity score. Here, we assume there is a single word w_e denoting e (e.g., *joy*).

3.2 Supervised Prediction

If a training set is available, then for each emotion e covered by it, we train a regression model $f_e(\mathbf{v}_w | \theta_e)$ parametrized by θ_e that allows us to predict the emotion intensity for e , given a vector of a word as input. Finally, we simply define

$$\sigma_s(w, e) = f_e(\mathbf{v}_w | \theta_e) \quad (2)$$

to route each prediction to the relevant f_e , i.e., given a target emotion e , we simply invoke the right model f_e .

In our experiments, we consider two kinds of models. The first is a feedforward neural network with a ReLU activated hidden layer with 64 neurons and a single output neuron that predicts the intensity of emotion, and a dropout rate of 0.2 applied to the hidden layer. The loss function is the mean squared error of the prediction with respect to the ground truth. The second option is a support vector regression (SVR) model with RBF kernel.

3.3 Self-Supervised Prediction

Finally, we propose a hybrid self-supervised technique that relies on supervised learning but does *not* require any pre-existing training data. Instead, we induce training data from the predictions of our unsupervised model. For each emotion e , we first identify the set T_e^+ , the words in the vocabulary \mathcal{V}

¹<http://emotion.nlproc.org/>

Method	Data	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Overall
Baselines	EmoLex	0.081	-0.071	-0.037	0.101	0.059	0.035	-0.186	0.250	0.076
	DepecheMood	0.089	–	–	0.077	-0.020	0.204	–	–	0.042*
	DepecheMood++	0.207	–	–	0.180	0.010	0.337	–	–	0.106*
	EmoWordNet	0.100	–	–	0.077	0.032	0.190	–	–	0.048*
Unsupervised	word2vec	0.321	0.266	0.056	0.081	0.482	0.321	0.249	0.403	0.254
	GloVe	0.360	0.413	0.031	0.137	0.352	0.281	-0.025	0.439	0.246
	fastText	0.388	0.332	0.071	0.038	0.306	0.281	0.186	0.497	0.249
	BERT	0.130	0.104	-0.136	0.081	0.002	-0.012	-0.037	0.155	0.033
	EWE	0.301	0.434	0.120	0.127	0.236	0.181	-0.143	0.348	0.194
	Counter-fitting	0.354	0.379	0.075	0.200	0.507	0.379	0.334	0.459	0.322
	AffectVec	0.426	0.430	0.262	0.447	0.631	0.559	-0.078	0.592	0.419
Self-sup. FFNN	word2vec	0.339	0.229	0.300	0.215	0.412	0.447	0.103	0.311	0.311
	GloVe	0.266	0.256	0.214	0.158	0.351	0.386	0.024	0.207	0.239
	fastText	0.352	0.300	0.348	0.272	0.347	0.530	0.137	0.469	0.359
	BERT	0.347	0.081	0.259	0.080	0.389	0.327	0.114	0.310	0.238
	EWE	0.259	0.242	0.251	0.174	0.330	0.304	-0.086	0.270	0.228
	Counter-fitting	0.345	0.305	0.177	0.205	0.467	0.353	0.119	0.445	0.300
	AffectVec	0.411	0.418	0.273	0.417	0.631	0.564	-0.091	0.574	0.405
Self-sup. SVR	word2vec	0.410	0.227	0.311	0.226	0.331	0.472	0.098	0.361	0.303
	GloVe	0.419	0.379	0.335	0.256	0.494	0.509	0.061	0.441	0.386
	fastText	0.346	0.302	0.409	0.272	0.389	0.513	0.158	0.480	0.371
	BERT	0.331	0.029	0.287	0.029	0.374	0.356	0.079	0.309	0.248
	EWE	0.348	0.327	0.386	0.249	0.390	0.486	-0.094	0.487	0.322
	Counter-fitting	0.368	0.349	0.224	0.213	0.471	0.388	0.138	0.469	0.314
	AffectVec	0.437	0.457	0.302	0.490	0.633	0.600	-0.128	0.605	0.426
Supervised FFNN	word2vec	0.681	0.568	0.711	0.713	0.695	0.749	0.712	0.694	0.699
	GloVe	0.704	0.559	0.733	0.723	0.678	0.731	0.643	0.730	0.698
	fastText	0.716	0.596	0.761	0.736	0.691	0.743	0.756	0.750	0.722
	BERT	0.624	0.486	0.550	0.578	0.578	0.570	0.612	0.580	0.533
	EWE	0.582	0.322	0.618	0.636	0.582	0.662	0.438	0.602	0.554
	Counter-fitting	0.675	0.580	0.641	0.676	0.660	0.649	0.732	0.693	0.665
	AffectVec	0.700	0.614	0.664	0.703	0.724	0.686	0.747	0.701	0.695
Supervised SVR	word2vec	0.642	0.566	0.654	0.654	0.680	0.734	0.721	0.664	0.672
	GloVe	0.720	0.633	0.727	0.671	0.719	0.723	0.722	0.694	0.705
	fastText	0.660	0.589	0.745	0.698	0.637	0.697	0.746	0.700	0.685
	BERT	0.567	0.478	0.484	0.532	0.576	0.559	0.683	0.556	0.553
	EWE	0.685	0.618	0.705	0.633	0.642	0.720	0.709	0.713	0.677
	Counter-fitting	0.654	0.569	0.558	0.623	0.618	0.636	0.711	0.665	0.630
	AffectVec	0.678	0.606	0.631	0.676	0.657	0.671	0.668	0.675	0.664

Table 1: Main emotion intensity prediction results, reported in terms of Pearson Correlation with the ground truth.

with the top k highest intensity predictions, and the set T_e^- , the words with the top k lowest intensity predictions.

$$T_e^+ = \operatorname{argmax}_{T \subset \mathcal{V}, |T|=k} \sum_{w \in T} \sigma_u(w, e) \quad (3)$$

$$T_e^- = \operatorname{argmin}_{T \subset \mathcal{V}, |T|=k} \sum_{w \in T} \sigma_u(w, e) \quad (4)$$

These predictions are acquired using Eq. 1, i.e., our unsupervised technique.

Subsequently, for each e , the set $T_e^+ \cup T_e^-$ along with the corresponding labels is used to train a supervised model $\bar{f}_e(\mathbf{v}_w | \bar{\theta}_e)$ as above in Section 3.2. Finally, the overall model again just involves invoking the relevant \bar{f}_e with regard to the desired emotion e :

$$\sigma_{\bar{s}}(w, e) = \bar{f}_e(\mathbf{v}_w | \bar{\theta}_e) \quad (5)$$

4 Evaluation

4.1 Main Results

We proceed with a detailed evaluation assessing to what extent a number of different linguistic resources and models capture emotion intensity information. For this, we consider the ground truth benchmark data introduced in Section 2 and compare the results obtained using different approaches with the test set ratings in terms of Pearson Correlation coefficients. The emotion intensity score for any out-of-vocabulary words is taken to be 0.0, indicating that the data does not make a prediction about such words.

The results are given in Table 1. The ‘‘Overall’’ column reports the correlation with the union of all word–emotion pairings in the ground truth test sets. As we have an equal number of word–emotion

pairs for each emotion, this serves as an aggregate measure of the result quality. For the baselines with fewer than eight emotions, the results are computed by averaging over just the subset of results for labels that are covered. A multi-way statistical significance analysis of the “Overall” results using t-tests shows that the improvements of all unsupervised, self-supervised, and supervised word vector methods except for unsupervised BERT are statistically significantly better than all lexicon baselines at significance level 0.001.

We compare a series of baselines along with our unsupervised, self-supervised, and regular supervised methods proposed in Section 3.

Baselines. We first evaluate existing state-of-the-art emotion lexicons. These provide either real-valued intensity scores or binary association scores (treated as 0.0 or 1.0). We found that the scores that the lexicons provide exhibit very low correlation with the ground truth. In the case of EmoLex (Mohammad and Turney, 2013), this is because it merely provides binary labels, not intensities.

For DepecheMood (Staiano and Guerini, 2014), DepecheMood++ (Araque et al., 2018), and EmoWordNet (Badaro et al., 2018), we conjecture that the data-driven automated techniques used to create them based on coarse-grained document-level labels do not result in word-level scores of the same sort as those solicited from human raters. The additional labels in DepecheMood and DepecheMood++ (e.g., Amused) may carry some information on some of the labels in our ground truth data (e.g., Joy). However, we find that mapping these emotions to a target emotion results in less accurate emotion association scores. Thus, we disregarded any labels not in the ground truth.

Unsupervised method. We next evaluate various pretrained word vector models, including word2vec trained on Google News (Mikolov et al., 2013), Glove 840B CommonCrawl (Pennington et al., 2014), and fastText trained on Wikipedia (Joulin et al., 2017). We find that these outperform the emotion lexicons by a substantial margin. They also outperform BERT, for which we use the pretrained BERT-base uncased model (Devlin et al., 2019) and consider mean-pooled word-piece final layer output embeddings as word-level vectors.

Even higher correlation can be attained with additional post-processing. One example are the counter-fitted vectors (Mrksic et al., 2016), ob-

tained by taking the PARAGRAM-SL999 vectors by Wieting et al. (2015) and optimizing them using synonymy and antonymy constraints.

The best results are obtained using AffectVec (Raji and de Melo, 2020), which post-processes the same vectors using not only synonymy and antonymy constraints, but also sentiment polarity ones. Sentiment polarity evidently helps to better distinguish different emotional associations.

In contrast, the emotion-enriched word vectors (EWE) by Agrawal et al. (2018) do not perform well for word intensity prediction.

Supervised method. For supervised methods, we report the mean correlation over 20 runs of the learning algorithm. The supervised models were chosen based on the proposed methods in the WASSA-2017 challenge (Mohammad and Bravo-Marquez, 2017), with minor simplifications to avoid overfitting, given the small size of our training set.

As expected, the models succeed in learning emotional associations from word vectors with a higher correlation than unsupervised prediction. As this technique does not rely on cosine similarities, post-processing proves fruitless, and very strong results are obtained using GloVe and fastText embeddings.

Self-supervised method. Unsupervised prediction is less accurate than supervised prediction, but is applicable to arbitrary emotions without any need for pre-existing emotion-specific training data. The self-supervised approach has the same advantage of not requiring training data and can thus as well be applied to arbitrary emotions.

To create the training labels for the self-supervised model, we select the top 100 ($k = 100$) highest and lowest intensity predictions from our unsupervised model. This results in 200 automatically induced training instances per emotion. For BERT, the vocabulary of fastText serves as the candidate set of words to calculate the top k predictions.

As expected, without access to gold standard training data, self-supervision is unable to compete with the supervised approach. However, for a given set of word vectors, especially for BERT, self-supervised learning mostly surpasses the unsupervised approach despite drawing on it for training, owing to its ability to selectively pick out the most pertinent cues from the vectors.

4.2 Unsupervised Sentence Classification

Experimental Setup. We additionally explore unsupervised *sentence-level* emotion classification on the GoEmotions dataset (Demszky et al., 2020) with its fine-grained inventory 28 different emotion labels. Given input document D , we choose the label

$$\arg \max_{e \in \Sigma} \sum_{w \in D} \gamma_{w,D} \sigma_u(e, w), \quad (6)$$

where Σ is the set of labels, $\gamma_{w,D}$ denotes the TF-IDF score of w in D , and $\sigma_u(w, e)$ are word-emotion scores. For the latter, we use either our unsupervised word scoring (Eq. 1) or a baseline lexicon, where any emotion not covered by a resource is assumed to have intensity scores of 0.0. An exception is made for the *neutral* label, which we choose if the average prediction score across labels in Σ is ≤ 0.03 , based on the reasoning that, similar to the annotation instructions in the original paper, low scores or conflicting ones should be labeled neutral.

Results. The evaluation results in Table 2 show that existing emotion lexicons are outperformed even by the random baseline of choosing a label arbitrarily. This is because they do not cover enough emotions to support this task—the GoEmotions dataset includes many fine-grained emotion labels such as *remorse*, *gratitude*, and *caring*. Emotion lexicons are often used for unsupervised analysis, but evidently can only work well when the target emotion inventory matches that of the lexicon.

In contrast, with several vector models, our unsupervised approach is able to greatly outperform all baselines, achieving substantially better results than chance. For reference, we also show the results of the fully *supervised* BERT model from Demszky et al. (2020), fine-tuned on the GoEmotions training data, while our vector-based models are entirely *unsupervised*. Note that the task is fairly challenging, since many of the 28 emotions are easy to confuse, e.g., *fear*, *nervousness*, *embarrassment*, *disappointment*, *disapproval*, etc.

5 Conclusion

In this paper, we show that pretrained linguistic models readily provide higher-quality emotional intensity information than state-of-the-art emotion lexicons, despite not being trained on emotion-labeled data, particularly if the representation space

Setup	Method	Precision	Recall	F ₁
Unsupervised	Random	3.6	3.6	3.6
	DepecheMood	1.7	5.2	1.8
	DepecheMood++	1.6	5.4	1.9
	EmoWordNet	1.7	5.1	1.8
	EmoLex	2.1	6.1	2.4
	GloVe	11.2	7.1	1.6
	fastText	24.6	9.5	4.9
	word2Vec	19.0	9.7	6.1
	Counter-fitting	15.6	16.4	10.0
	AffectVec	18.3	15.5	10.1
	Supervised	Fine-Tuned BERT	40	63

Table 2: Results for 28-way unsupervised sentence classification on GoEmotions dataset, in terms of macro-averaged precision, recall, and F₁-score in %.

is post-processed. We find that a regular supervised variant obtains very high correlations, while a self-supervised variant that does not require gold standard training data is able to outperform the unsupervised method and can as well be applied to arbitrary emotion labels. Overall, our results confirm that linguistic models have remained underexplored for word-level emotion intensity assessment. Our training/validation/test splits and newly induced lexicons created using our method are available from <http://emotion.nlproc.org/> to promote further research in this area.

Broader Impact

Assessing the emotions evoked by a text has important applications, such as discovering disappointed customers on social media, designing conversational agents that emulate human empathy by responding in a more appropriate way, as well as numerous forms of digital humanities analyses. However, there are also grave concerns and risks when using automated techniques to assess the emotional impact of text. Clearly, consent to assess the text must have been granted and malicious as well as potentially harmful applications must be avoided.

Beyond this, even for applications deemed beneficial, there is a risk that inaccurate assessments may lead to undesirable outcomes. The empirical results in this paper show that many commonly used resources and techniques have a low correlation with human ratings and even the best BERT-based learning approaches are highly imperfect, as they may latch on to superficial lexical cues and mere correlations that may exhibit substantial biases. This is discussed further by Mohammad (2020). As a result, prior to any potential use of such techniques, a thorough analysis of potential application-specific risks must be conducted.

References

- Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. FEEL: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.
- Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. [Learning emotion-enriched word representations](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA.
- Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2018. DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. *arXiv preprint arXiv:1810.03660*.
- Gilbert Badaro, Hussein Jundi, Hazem Hajj, and Wasim El-Hajj. 2018. [EmoWordNet: Automatic expansion of emotion lexicon using English WordNet](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 86–93, New Orleans, Louisiana.
- Sven Buechel and Udo Hahn. 2018a. [Emotion representation mapping for automatic lexicon construction \(mostly\) performs on human level](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904, Santa Fe, New Mexico, USA.
- Sven Buechel and Udo Hahn. 2018b. [Word emotion induction for multiple languages as a deep multi-task learning problem](#). 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 1907–1918, New Orleans, Louisiana.
- Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle H. Ungar. 2018. [Learning neural emotion analysis from 100 observations: The surprising effectiveness of pre-trained word representations](#). *CoRR*, abs/1810.10949.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). *CoRR*, abs/2005.00547.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. [Placing search in context: The concept revisited](#). In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 406–414, New York, NY, USA. ACM.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2014. [Simlex-999: Evaluating semantic models with \(genuine\) similarity estimation](#). *CoRR*, abs/1408.3456.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.
- Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. [Aff2vec: Affect-enriched distributional word representations](#). *CoRR*, abs/1805.07966.
- Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70.
- 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 26*, 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, Copenhagen, Denmark.
- Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.
- Saif M. Mohammad. 2020. [Practical and ethical considerations in the effective use of emotion and sentiment lexicons](#).
- Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.
- Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. [Counter-fitting word vectors to linguistic constraints](#). *CoRR*, abs/1603.00892.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Robert Plutchik. 1980. [A general psychoevolutionary theory of emotion](#). In *Theories of Emotion*, pages 3–33. Elsevier.
- Shahab Raji and Gerard de Melo. 2020. What sparks joy: The AffectVec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997, New York, NY, USA. ACM.

- Arun Ramachandran and Gerard de Melo. 2020. [Cross-lingual emotion lexicon induction using representation alignment in low-resource settings](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5879–5890, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. [Ultradense word embeddings by orthogonal transformation](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Herbert Rubenstein and John B. Goodenough. 1965. [Contextual correlates of synonymy](#). *Commun. ACM*, 8(10):627–633.
- Abu Awal Md Shoeb and Gerard de Melo. 2020. [EmoTag1200: Understanding the association between emojis and emotions](#). In *Proceedings of EMNLP 2020*, pages 8957–8967. Association for Computational Linguistics.
- Abu Awal Md Shoeb, Shahab Raji, and Gerard de Melo. 2019. [EmoTag – Towards an emotion-based analysis of emojis](#). In *Proceedings of RANLP 2019*, pages 1094–1103.
- Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Trevino, and Juan Gordon. 2012. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence*, pages 1–14. Springer.
- 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.
- Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. [From paraphrase database to compositional paraphrase model and back](#). *Transactions of the Association for Computational Linguistics*, 3:345–358.