IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning

Maximilian Köper, Evgeny Kim and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,evgeny.kim,roman.klinger}@ims.uni-stuttgart.de

Abstract

Our submission to the WASSA-2017 shared task on the prediction of emotion intensity in tweets is a supervised learning method with extended lexicons of affective norms. We combine three main information sources in a random forrest regressor, namely (1), manually created resources, (2) automatically extended lexicons, and (3) the output of a neural network (CNN-LSTM) for sentence regression. All three feature sets perform similarly well in isolation ($\approx$ .67 macro average Pearson correlation). The combination achieves .72 on the official test set (ranked 2nd out of 22 participants). Our analysis reveals that performance is increased by providing cross-emotional intensity predictions. The automatic extension of lexicon features benefit from domain specific embeddings. Complementary ratings for affective norms increase the impact of lexicon features. Our resources (ratings for 1.6 million twitter specific words) and our implementation is publicly available at http://www.ims.uni-stuttgart.de/data/ims_emoint.

1 Introduction

In natural language processing, emotion recognition is the task of associating words, phrases or documents with predefined emotions from psychological models. Typical discrete categories are those proposed by Ekman (Ekman, 1999) and Plutchik (Plutchik, 2001), namely *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise* und *Trust*. In contrast to sentiment analysis with its main task to recognize the polarity of text (*e. g.*, positive, negative, neutral, mixed), only a few resources and domains have been subject of analysis. Examples are, *e. g.*, tales (Alm et al., 2005), blogs (Aman and Szpakowicz, 2007), and as a very popular domain, microblogs on Twitter (Dodds et al., 2011). The latter in particular provides a large resource of data in the form of user messages (Costa et al., 2014). A common source of weak supervision for training classifiers are hashtags, emoticons, or emojis, which are interpreted as a weak form of author "self-labeling" (Suttles and Ide, 2013). The classifier then learns the association of all other words in the message with the emotion (Wang et al., 2012). An alternative to discrete models are continuous models that map emotions to an n-dimensional space with valence, arousal and dominance (VAD) being usual dimensions. Previous works that rely on the VAD-scheme focus mainly on extending and adapting the affective lexicons (Bestgen and Vincze, 2012; Turney and Littman, 2003), including to historical texts (Buechel et al., 2016), and on the prediction and extrapolation of affective ratings (Recchia and Louwerse, 2015a; Hollis et al., 2017).

The WASSA-2017 shared task on the prediction of emotion intensity in tweets (EmoInt) aims at combining descrete emotion classes with different levels of activation. Given a tweet and an emotion (*anger*, *fear*, *joy*, and *sadness*), the task requires to determine the intensity expressed regarding a particular emotion. This score can be seen as an approximation of the emotion intensity felt by the reader or expressed by the author. For a detailed task descriptions and background information on the data collection see Mohammad and Bravo-Marquez (2017).

2 System Description

In the following, we introduce all feature sets we experimented with. We start with an analysis and selection of features obtained from the baseline

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 50–57
Copenhagen, Denmark, September 7–11, 2017.

Rating	Top 4 words
Concreteness	fish, microphone, rope, toilet
Arousal	#attack, scare, attack, exciting
Dominance	#safe, #everydayhappy, courageous, #Amoved
Happiness	babygiggles, love, laughter, lovelysmile
Anger	soangry, comcastsucks, #soangry, #comcastsucks
Fear	#hyperventilation, #irrationalfear, aerophobia, #anxiety
Sadness	#greatloss, greatloss, sadsadsad, cryinggame
Joy	#peaceandharmony, #alwaysbethankful, positiveenergy, #youchoosehowtofeel

Table 1: Top four words for eight different rating types based on our automatically generated ratings.

system AffectiveTweets, explain how we extend resources to the domain of Twitter. Then, we explain our sentence regressor, which is based on deep learning and pre-trained word embeddings. Finally, we introduce two additional, manually defined features.

2.1 Baseline Features

The baseline system *AffectiveTweets*[1] which has been provided to participants together with the training and development data includes a huge variety of different features and configurations. The different feature types can be classified into a), *SparseFeatures*, which refer to word and character n-grams from tweets, b), *LexiconFeatures*, which are taken from several emotion and sentiment lists (we consider the *SentiStrength*-based feature to be part of this), and c), the *EmbeddingsFeature*, which comprise a tweet-level feature representation that can incorporate any pre-trained word embeddings.

2.2 Extending and Adding Norms

The baseline system builds on top of a variety of different lexical resources (Hu and Liu, 2004; Wilson et al., 2005; Svetlana Kiritchenko and Mohammad; Mohammad and Turney, 2013; Mohammad and Kiritchenko, 2015; Baccianella et al., 2010; Bravo-Marquez et al., 2016; Nielsen, 2011). Such resources are naturally limited in coverage and often focus on words that are closely associated with a certain emotion or sentiment (*e. g.*, the word "hate" with the emotion *anger*).

At the same time, social media data is typically rich in lexical variations, and hence, tend to contain a great deal of out-of-vocabulary words. We address this with three separate approaches, namely by i) applying a supervised method to extend these lexicons to larger Twitter specific vocabulary ii), learning a new rating score for every word and not just highly associated terms and iii), including novel rating categories that provide complementary and potential useful information, such as valence, arousal, dominance and concreteness.

Several approaches have been proposed to combine distributional word representations with supervised machine learning methods to extend affective norms (Turney et al., 2011; Tsvetkov et al., 2014; Recchia and Louwerse, 2015b; Vankrunkelsven et al., 2015; Köper and Schulte im Walde, 2016; Sedoc et al., 2017). Köper and Schulte im Walde (2017) compared various supervised methods and showed that a feed forward neural network together with low dimensional distributed word representations (embeddings) obtained the highest correlation with human annotated ratings for concreteness.

Following these findings, we apply the same methodology. For a given emotion or norm we train a feed forward neural network with two hidden layers, each having 200 neurons. The input of the network is a single word representation (300 dimensions) and the output is one numerical value trained to correspond to the human annotated (gold) rating for the given input word. We apply the model to predict a rating score for every word representation in our distributional space (which includes the training data).

This method strongly depends on the underlying word representation. We therefore conduct multiple experiments using different word embeddings (shown in Section 4.2). We apply this procedure for 13 different lexicons using the following resources: *NRC Hashtag Emotion Lexicon* (Mohammad and Kiritchenko, 2015) containing ratings for $17k$ words with associations to *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. Additionally, we use the $14k$ ratings for *valence*, *arousal*, and *dominance* collected by Warriner et al. (2013). For *concreteness* we rely on the collection of $40k$ ratings from Brysbaert et al. (2014). Finally,

[1] https://github.com/felipebravom/AffectiveTweets

we use the $10k$ ratings for *happiness* from Dodds et al. (2011). These 13 ratings correspond to an automatic extension to 1.6 million word types with ≈ 21 million new word ratings. We map the ratings to an interval of $[0, 10]$. Table 1 shows the top words for eight ratings. For the emotion intensity prediction in our predictive model, we represent each rating with seven feature dimensions per tweet:

1. Average rating score across all words
2. Average rating score across all nouns
3. Average rating score across all adjectives
4. Average rating score across all verbs
5. Average rating score across all hashtags
6. Maximum rating score
7. Standard deviation of all rating scores

2.3 Tweet Regression

The tweet regression feature relies on the annotated training samples. We train a neural network based on word embeddings to predict the emotion intensity for each tweet.

Convolutional neural networks (CNNs), trained on top of pre-trained word vectors, have been shown to work well for sentence-level classification tasks (Kim, 2014). We apply a similar method here, combining CNNs and LSTMs (Hochreiter and Schmidhuber, 1997). The final architecture used by `IMS` is shown in Figure 1. Each tweet is represented by a matrix of size 50×300 (padded where necessary, embedding dimension is 300, the maximal token sequence in a tweet is set to 50). We apply dropout with a rate of 0.25. The matrix is then the input for a convolutional layer with a window size of 3, followed by a maxpooling layer (size 2) and an LSTM to predict a numerical output for each tweet.

This architecture captures sequential information in a compact way. For comparison, we conduct experiments using a variety of different architectures (shown in Section 4.3) including linear regression, multilayer perceptron (MLP), two stacked LSTMs and the proposed CNN-LSTM architecture.

2.4 Additional Features

In addition to regression and lexical features, we add two hand-crafted features. The first is a Boolean feature which holds if and only if an exclamation mark is present in the tweet. The second represents the overall number of tokens in the tweet.

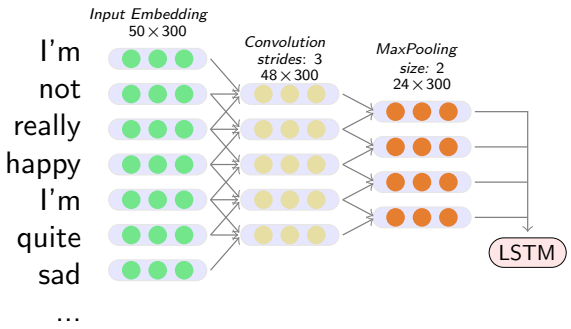

Figure 1: CNN-LSTM Architecture used for tweet regression.

3 Implementation Details

As a source for our in-domain embeddings, we use a corpus from 2016 retrieved with the Twitter streaming and rest APIs with emotion hashtags and popular general hashtags. It consists of ≈50 million tweets and ≈800 million tokens. After removing words with less than 10 occurrences, the resource contains 1.6 million word types. The 300 dimensional word representations are obtained with *word2vec*[2] (Mikolov et al., 2013). To study the impact of the training domain, we additionally conduct experiments with the public available *GoogleNews-vectors* that were trained on a 100b words corpus of news texts. Both word embeddings are used to extend the emotion lexicons (Section 2.2) as well as input embeddings in our tweet regression model (Section 2.3).

We use *TweetNLP*[3] (Owoputi et al., 2013) as tokenizer. In the case of observing only out-of-vocabulary words (no rating available) we set the score to the median value of the corresponding category.

The regressor based on the tweet text is implemented with *keras* (Chollet et al., 2015). We train one model for each of the four emotions separately. Furthermore, we provide the output of all four emotion-specific regression models in all emotion intensity prediction tasks.[4]

Finally, for the full system `IMS`, we combine features in a random forest classifier using *weka* (Witten et al., 1999). We use 800 trees (called *iterations* in Weka). We estimate one model for each of the four target emotions.

[2]Hyperparameters were set to window:5, min-count:10, neg-samples:15, dim:300, iteration:5.

[3]`http://www.cs.cmu.edu/~ark/TweetNLP/`

[4]To provide this feature for the within-emotion training data (*e. g.*, anger-regression output for anger training dataset), we split the training data into 20 folds – training on 19 and

Feature	Model	a	f	j	s	Avg
✓ Lexicons	✗ SVM	.62	.62	.62	.62	.62
	✓ RF	**.67**	**.69**	**.66**	**.66**	**.67**
✗ Sparse	✗ SVM	.58	.61	.63	.52	.58
	✗ RF	.53	.57	.61	.53	.56
✗ Embd.	✗ SVM	.48	.50	.55	.53	.51
	✗ RF	.53	.53	.61	.49	.54
✗ Comb	✗ SVM	.64	.64	.66	.64	.64
	✗ RF	.63	.64	.66	.63	.64

Table 2: Baseline features across training data using support vector machines (SVM) and random forest (RF). Pearson correlation based on 10-fold cross validation. The column names denote anger (a), fear (f), joy (j), sadness (s).

Feat	a	f	j	s	Avg
✓ Lexicons(=BL)	**.67**	.69	**.66**	.66	**.67**
✗ ACVH-Lexicons	.48	.45	.59	.35	.47
✗ Ext.News	.52	.52	.60	.44	.52
✓ Ext.Twitter	.65	**.69**	.65	**.68**	.67
✗ ACVH-Lexicons+BL	.66	.67	.67	.64	.66
✗ Ext.News+BL	.65	.66	.67	.64	.65
✓ Ext.Twitter+BL	**.68**	**.71**	**.68**	**.69**	**.69**

Table 3: Performance of lexicons and our automatically extended lexicons. Results are based on the random forest classifier. Top part compares performance of lexicon features in isolation. Ext.News and Ext.Twitter build on top of the baseline lexicons and the ACVH lexicons. The bottom part shows performance in combination with the original lexicons provided by the baseline (=BL).

4 Feature Subset Selection and Analysis

Feature selection and analysis was performed on annotated training and development data. All experiments were carried out using 10-fold cross validation. We report results following the official shared task evaluation measure to predict a value between 0 and 1, namely Pearson correlation for each emotion separately as well as a macro average over all emotions. Features that were finally used in `IMS` are marked with ✓ and respectively ✗ for features that were disregarded.

4.1 Baseline Feature Engineering

We start with feature engineering based solely on the baseline features (see Section 2.1). Table 2 shows our observation when exploring the different options from *AffectiveTweets* using default parameters. The embeddings (Embd.) are the recommended 400 dimensional Twitter embeddings available from the baseline system's homepage.

As we see in this table, an average performance of .67 is already obtained when relying only on a random forest in combination with the lexicon features. The other features, as well as the combination, result in inferior performance. In addition, the lexicon-based system is comparably simple with only 45 feature dimensions. We therefore only use the lexicon features from the baseline system.

4.2 Lexicons and Extended Lexicons

As a next feature, we explore various settings for the automatic extension of the lexicon features. Table 3 compares the baseline lexicon against the lexicons we add without extension (*ACVH-Lexicons*) as well as the automatically extended resources (*Ext.**). *ACVH-Lexicons* contains the unmodified ratings for arousal, concreteness, valency and happiness (ACVH), which were not part of the baseline system. For *Ext.** we present results based on underlying news (*Ext.News*) and Twitter (*Ext.Twitter*) embeddings. In addition we present results for each lexicon-feature in isolation, as well as in combination with the baseline lexicons (*Lexicons(=BL)*). It can be seen that the ACVH lexicons without automatic extension (*ACVH-Lexicons*) perform poorly and provide no performance gain when combined with the baseline (*ACVH-Lexicons+BL*). We assume that the poor coverage on Twitter data is the main reason. On the other hand, the automatically extended ratings perform well, and the choice of embeddings here has a high impact on the quality of the resulting ratings. In more detail, the in-domain embeddings (*Ext.Twitter*) create ratings that are extrinsically evaluated superior to the out-domain embeddings (*Ext.News*) with an average score .52 against .67.

The information of existing lexicons and extended norms is not redundant. The combination (*Ext.Twitter+BL*) increases average correlation across all four emotions by $+.02$ points, from .67 $\rightarrow$ .69..

To get a further understanding of the automatically extended norms, Figure 2 shows the evaluation performance of the thirteen extended norm

providing the predictions for the remaining.

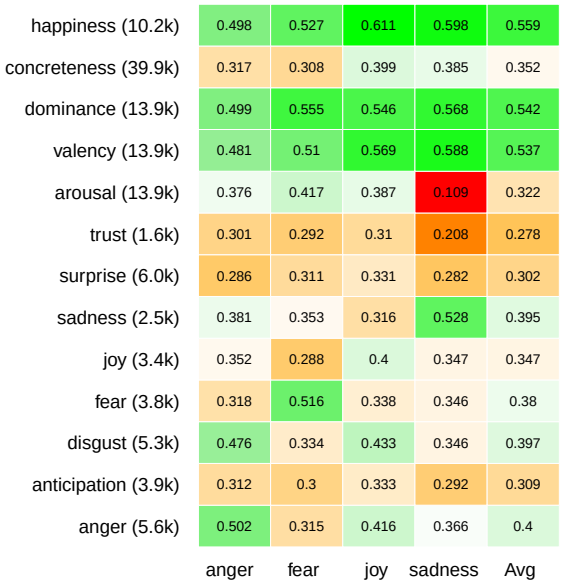

Figure 2: Pearson's correlation of single rating categories (Y-Axis) on each target emotion (X-Axis). Numbers in brackets refer to training size used to extend the norms. Evaluation based on 10-fold cross validation using the full training data and random forest.

Feature	a	f	j	s	Avg
✗ Linear Reg. (BoW)	.48	.49	.44	.36	.44
✗ MLP (BoW)	.59	.64	.60	.56	.60
✗ Stacked LSTMs	.58	.66	.61	.61	.61
✓ CNN-LSTM	**.66**	**.68**	**.66**	**.65**	**.67**

Table 4: Comparing the performance of Tweet Regression Architectures.

categories separately. Especially the extended ratings from the new lexicons perform well: *happiness*, *dominance* and *valency*. However, we also see that the number of training samples might have a big impact, *e. g.*, the automatical ratings of joy are only trained on $3.4k$ samples while the size of the *happiness* training data is larger.

4.3 Tweet Regression Architectures

In addition to the CNN-LSTM architecture used in the final system (see Section 2.3), we experimented with different models for tweet regression. Table 4 shows results using various machine learning algorithms to directly predict the emotion intensity.

We use the in-domain Twitter embeddings as input. We observe that our architecture, introduced in Section 2.3, performs superior to other methods. Remarkable, the CNN-LSTM feature, as well as our *Ext.Twitter* lexicons and the baseline *Lexicons(=BL)* obtain a score of $\approx$ .66 when used in isolation.

Feature Name	# Features
AffectiveTweets-Lexicons	45
Aut. Ext. Lexicons (Twitter)	91
Tweet Regression (CNN-LSTM)	4
Manual Features	2
Total	142

Table 5: Overview `IMS` full system, features, feature counts.

Full `IMS`-Train	**a**	**f**	**j**	**s**	**Avg**
	.71	.74	.71	.71	.72

Table 6: Final official system on training data (10 fold cross validation).

4.4 Full System Combination

A combination of all features leads to the best performance, they provide complementary information. An overview is given in Table 5 and Table 6.

Another interesting observation is found with respect to the usage of cross-emotional intensity predictions: `IMS` trains a classifier for each emotion in isolation. Similarly, the tweet regression feature is trained emotion-wise but for each instance we also provide the intensity prediction from all other emotion models (therefore, 4 features). Without the cross-emotion information, we yield only a macro average across all emotions of .707 (vs. .719). Figure 3 shows how the emotion intensity predictions of these models correlate. It can be seen that *fear*, *sadness* and *anger* are slightly correlated while *joy* is negatively correlated with all three emotions. Interestingly, a combined model (*Comb*), which is trained on all emotions also leads to a high correlation for each emotion and especially *sadness*. Note that the classifier trained on all emotions (*Comb*) is not used by the final system `IMS`.

Finally, we want to mention that the impact of the two manual defined features is very little, we found that they increase performance on *joy* by $+.01$ and we therefore decided to keep them.

5 Official Results – Analysis Test Data

Table 7 shows the official results (Full `IMS`-Test) and the performance using only a subset of the

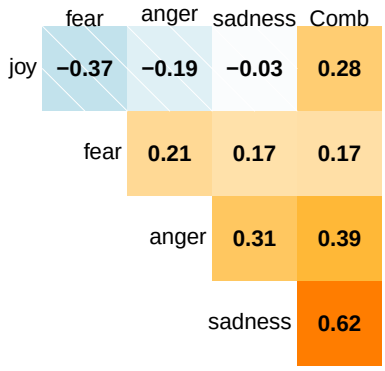

Figure 3: Pairwise Pearson correlation based on the output of our emotion-wise Tweet regression feature.

Feat	a	f	j	s	Avg
Lexicons(=BL)	.65	.66	.60	.70	.65
Ext.Twitter+BL	.68	.72	.66	.74	.70
CNN-LSTM+BL	.69	.69	.67	.76	.70
Full IMS-Test	.71	.73	.69	.77	.72
Best-Competitor	.73	.76	.73	.76	.75

Table 7: Overview IMS full and partial System performance on Test data.

entire features. For comparison, we also show the results of the best performing system (Best-Competitor). our baseline, using only the lexicon features and a random forest classifier obtains a competitive Pearson correlation of .65, which would have been ranked as the 8th best system.

Both of our core features, namely the extended resources, as well as the CNN-LSTM tweet regression architecture, increase performance by $+.05$ points when combined with the baseline lexicons (Lexicons(=BL)). Their performance is similar for *anger* and *joy*, but the ratings seem more useful for *fear*, and the regression more useful for *sadness*. The result of *Ext.Twitter+BL* with .70 would have ranked the 4th best system.

The final combination of all our features results in an increase of $\approx +.020$ correlation points. The performance of IMS on the test set without the two manually defined features is .719. Furthermore, we observe that our submission on the test data is on average very close to the estimated performance on the training data (both .72), but when looking at individual emotions our system is performing better on *sadness* and slightly worse on *fear*.

5.1 Error Analysis

Based on a manual inspection of individual tweets with a large gap between prediction and gold rating, we found that the model's prediction often depends on single words and ignores larger contexts. An example case with a high error for *fear* is:

> *"Most people never achieve their goals because they are afraid to fail."*
> *(fear, G: .22, P: .55)*

Here, the gold emotion intensity for *fear* is comparably low, but our model predicts a high fear intensity. Similarly, in the tweet with high joy intensity

> *"Just died from laughter after seeing that."*
> *(joy, G: .92, P: .50)*

our model predicts a low joy intensity.

Another challenge are modifications as in

> *"After this news Im supposed to be so damn happy and rejoicing but Im here like* □*"*
> *(joy, G: .07, P: .53)*

Here, the gold annotation is very low, but our model predicts a medium intensity for joy.

6 Conclusion

Our system IMS, submitted to the *EmoInt-2017* shared task, combines existing lexicons with automatically extended norms and a CNN-LSTM neural network based on embeddings. Our findings show that each of the three main components performs equally well, but the highest performance is achieved in combination. In addition, we found that extending existing emotion lexicons and affective norms improves performance over the original resources. We also showed that the impact of underlying word representation is important. In particular in-domain embeddings (trained on twitter data) perform superior to other embeddings. A particularly interesting observation is that providing cross-emotional intensity predictions benefits the performance.

7 Acknowledgement

The research was supported by the DFG Collaborative Research Centre SFB 732 and the German Ministry for Education and Research (BMBF) within the Center for Reflected Text Analytics (CRETA). We thank the anonymous reviewers for their comments and Jeremy Barnes for helpful suggestions.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *Proceedings of HLT-EMNLP*, pages 579–586, Vancouver, BC.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of TSD*, pages 196–205, Plzeň, Czech Republic.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Yves Bestgen and Nadja Vincze. 2012. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods*, 44(4):998–1006.

Felipe Bravo-Marquez, Eibe Frank, Saif M. Mohammad, and Bernhard Pfahringer. 2016. Determining Word-Emotion Associations from Tweets by Multi-label Classification. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016*, pages 536–539.

Marc Brysbaert, AmyBeth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally known English Word Lemmas. *Behavior Research Methods*, pages 904–911.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the PastAdapting Affective Lexicons for Historical Emotion Analysis. *LT4DH 2016*, page 54.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

Joana Costa, Catarina Silva, Mário Antunes, and Bernardete Ribeiro. 2014. Concept Drift Awareness in Twitter Streams. In *Proceedings of ICMLA*, pages 294–299, Detroit, MI.

Peter Sheridan Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss, and C. M. Danforth. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. Draft version. Available at http://arxiv.org/abs/1101.5120v3. Accessed October 24, 2011.

Paul Ekman. 1999. Basic emotions. In M Dalgleish, T; Power, editor, *Handbook of Cognition and Emotion*. John Wiley & Sons, Sussex, UK.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Geoff Hollis, Chris Westbury, and Lianne Lefsrud. 2017. Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, 70(8):1603–1619.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portoro, Slovenia.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving Verb Metaphor Detection by Propagating Abstractness to Words, Phrases and Individual Senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain.

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 *In Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, Copenhagen, Denmark.

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. 29(3):436–465.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*, pages 93–98.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *In Proceedings of NAACL*, pages 380–390, Atlanta,GA, USA.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(July–August):344–350.

Gabriel Recchia and Max M Louwerse. 2015a. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68(8):1584–1598.

Gabriel Recchia and Max M. Louwerse. 2015b. Reproducing Affective Norms with Lexical Co-occurrence Statistics: Predicting Valence, Arousal, and Dominance. *The Quarterly Journal of Experimental Psychology*, 68(8):1584–1598.

Joao Sedoc, Daniel Preotiuc-Pietro, and Lyle Ungar. 2017. Predicting Emotional Word Ratings using Distributional Representations and Signed Clustering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, Valencia, Spain.

Jared Suttles and Nancy Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. In *Proceedings of CiCLing*, volume 7817 of *Lecture Notes in Computer Science*, pages 121–136. Springer.

Xiaodan Zhu Svetlana Kiritchenko and Saif M. Mohammad. Sentiment Analysis of Short Informal Texts. 50:723–762.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Peter D Turney and Michael L Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Hendrik Vankrunkelsven, Steven Verheyen, Simon De Deyne, and Gerrit Storms. 2015. Predicting Lexical Norms Using a Word Association Corpus. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, California, USA.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *Proceedings of SocialCom/PASSAT*, pages 587–592, Amsterdam, Netherlands.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 347–354.

Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical Machine Learning Tools and Techniques with Java Implementations.