# Modelling Valence and Arousal in Facebook posts

**Daniel Preoţiuc-Pietro**
Positive Psychology Center
University of Pennsylvania
danielpr@sas.upenn.edu

**H. Andrew Schwartz**
Department of Computer Science
Stony Brook University
has@cs.stonybrook.edu

**Gregory Park** and **Johannes C. Eichstaedt**
Positive Psychology Center
University of Pennsylvania

**Margaret Kern**
Centre for Positive Psychology
University of Melbourne

**Lyle Ungar**
Computer & Information Science
University of Pennsylvania
ungar@cis.upenn.edu

**Elizabeth P. Shulman**
Department of Psychology
Brock University
eshulman@brocku.ca

## Abstract

Access to expressions of subjective personal posts increased with the popularity of Social Media. However, most of the work in sentiment analysis focuses on predicting only valence from text and usually targeted at a product, rather than affective states. In this paper, we introduce a new data set of 2895 Social Media posts rated by two psychologically-trained annotators on two separate ordinal nine-point scales. These scales represent valence (or sentiment) and arousal (or intensity), which defines each post's position on the circumplex model of affect, a well-established system for describing emotional states (Russell, 1980; Posner et al., 2005). The data set is used to train prediction models for each of the two dimensions from text which achieve high predictive accuracy – correlated at $r = .65$ with valence and $r = .85$ with arousal annotations. Our data set offers a building block to a deeper study of personal affect as expressed in social media. This can be used in applications such as mental illness detection or in automated large-scale psychological studies.

## 1 Introduction

Sentiment analysis is a very active research area that aims to identify, extract and analyze subjective information from text (Pang and Lee, 2008). This generally includes identifying if a piece of text is subjective or objective, what sentiment it expresses (positive or negative; often referred to as valence), what emotion it conveys (Strapparava and Mihalcea, 2007) and towards which entity or aspect of the text i.e., aspect based sentiment analysis (Brody and Elhadad, 2010). Downstream applications are mostly interested in automatically inferring public opinion about products or actions. Besides expressing attitudes towards other objects, texts can also express the emotions of the ones writing them, most common recently with the rise of Social Media usage (Rosenthal et al., 2015). This study focuses on presenting a gold standard data set as well as a model trained on this data in order to drive research in learning about the affective norms of people posting subjective messages. This is of great interest to applications in social science which study text at a large scale and with orders of magnitude more users than traditional studies.

Emotion classification is a widely debated topic in psychology (Gendron and Barrett, 2009). Two main theories about emotions exist: the first posits a discrete and finite set of emotions, while the second suggests that emotions are a combination of different scales. Research in Natural Language Processing (NLP) has been focused mostly on Ekman's model of emotion (Ekman, 1992) which posits the existence of six basic emotions: anger, disgust, fear, joy, sadness and surprise (Strapparava and Valitutti, 2004; Strapparava and Mihalcea, 2008; Calvo and D'Mello, 2010). In this study, we focus on the most popular dimensional model of emotion: the circumplex model introduced in (Russell, 1980). This model suggests that all affective

states are represented in a two-dimensional space with two independent neurophysiological systems: valence (or sentiment) and arousal. Any affective experience is a linear combination of these two independent systems, which is then interpreted as representing a particular emotion. For example, fear is a state involving the combination of negative valence and high arousal (Posner et al., 2005). Previous research in NLP focused mostly on valence or sentiment, either binary or having a strength component coupled with sentiment (Wilson et al., 2005; Thelwall et al., 2010; Thelwall et al., 2012).

In this paper we build a new data set consisting of 2895 anonymized Facebook posts labeled with both valence and arousal by two annotators with psychology training. The ratings are made on two independent nine point scales, reaching a high agreement correlations of .768 for valence and .827 for arousal. Data set statistics suggest that while the dimensions of valence and arousal are associated, they present distinct information, especially in posts with a clear positive or negative valence.

Further, we train a bag-of-words linear regression model to predict ratings of new messages. This model achieves high correlation with actual mean ratings, reaching Pearson $r = .85$ correlation on the arousal dimension and $r = .65$ on the valence dimension without using any other sentiment analysis resources. Comparing our method to other established lexicons for valence and arousal and methods from sentiment analysis, we demonstrate that these methods are not able to handle well the type of posts present in our data set. We further illustrate the most correlated words with both dimensions and identify opportunities for improvement. The data set and annotations are freely available online.[1]

## 2 Data set

We create a new data set with annotations on two independent scales:

- **Valence (or sentiment)** represents the polarity of the affective content t in a post, rated on a nine point scale from 1 (very negative) to 5 (neutral/objective) to 9 (very positive);
- **Arousal (or intensity)** represents the intensity of the affective content, rated on a nine point scale from 1 (neutral/objective post) to 9 (very high).

Our corpus is comprised of Facebook status updates shared by participants as part of the MyPersonality Facebook application (Kosinski et al., 2013), in which they also took a variety of questionnaires. All authors have explicitly given permission to include their information in a corpus for research purposes. We have manually anonymized the entire corpus by removing any references to other names of persons, addresses, telephone numbers, emails and URLs, and replaced them with placeholders.

In order to reduce biases due our participant demographics, the data set sample was stratified by gender and age and we have not rated more than two messages written by the same person. Research is inconclusive about whether females express more emotions in general (Wester et al., 2002). With regards to age, an age positivity bias has been found, where positive emotion expression increases with age (Mather and Carstensen, 2005; Kern et al., 2014).

The data originally consisted of 3120 posts. All of these posts were annotated by the same two independent raters with a training in psychology. The raters performed the coding in a similar environment without any distractions (e.g., no listening to music, no watching TV/videos) as these could have influenced the emotions of raters, and therefore the coding.

The annotators were instructed to sparingly rate messages as *un-ratable* when they were written in other languages than English or that offered no cues for a accurate rating (only characters with no meaning). The annotators were instructed to rate a message if they could judge at least a part of the message. Then, the raters were asked to rate the two dimensions, valence and arousal, after they have explicitly been briefed that these should be independent of each other. The raters were provided with anchors with specified valence and arousal and were instructed to rate neutral messages at the middle of the scale in terms of valence and 1 if they lacked arousal.

[1] http://mypersonality.org/wiki/doku.php?id=download_databases

| Dimension | R1 $\mu \pm \sigma$ | R2 $\mu \pm \sigma$ | IA Corr. |
|---|---|---|---|
| Valence | $5.274 \pm 1.041$ | $5.250 \pm 1.485$ | .768 |
| Arousal | $3.363 \pm 1.958$ | $3.342 \pm 2.183$ | .827 |

Table 2: Individual rater mean and standard deviation and inter-annotator correlation (IA Corr).

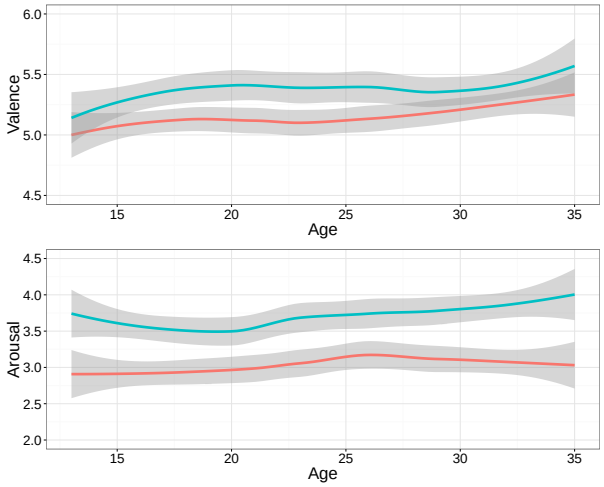



Figure 2: Variation in valence and arousal with age in our data set using a LOESS fit. Data is split by gender: Male (coral orange) and Female (mint green).

In total, 2895 messages were rated by both users in both dimensions. Table 1 shows examples of posts rated in all quadrants of the circumplex model.

The correlation between the raters and the mean and standard deviation for each rater are presented in Table 2. The inter-annotator agreement on deciding un-ratable posts is measured by Cohen's Kappa of $\kappa = .93$. The histograms of ratings are presented in Figure 1. The data set is released with the scores of both individual raters.

We study the correlation between the valence and arousal scores for posts in Table 3. We chose to split values based on different valence thresholds in order to remove posts rated as neutral in valence (5) from the analysis, as they are expected to be low in intensity (1). We observed an overall correlation between the valence and arousal ratings, which holds for both positive and negative valence tweets when the neutral posts are removed (.222, .226 correlation). However, when the posts are both more positive and negative in valence, arousal is only mildly correlated (.047 and .085). This highlights that the

| Valence of posts | 1–9 | 1–3.5 | 1–4 | 6–9 | 6.5–9 |
|---|---|---|---|---|---|
| Correlation to arousal | .222 | -.047 | -.201 | .226 | .085 |
| Mean arousal | 3.35 | 3.85 | 3.47 | 4.31 | 4.68 |

Table 3: Correlation with arousal and mean arousal values for different posts grouped by valence.

presence of either positive and negative valence is correlated with a arousal score different than 1, but this correlation is weaker when the positive or negative valence passes a certain threshold (i.e. 3.5 and 6.5 respectively). We also note that the high overall correlation is also due to higher mean arousal for positive valence posts compared to negative posts (4.68 cf. 3.85)

Figure 2 displays the relationship between the age of the user at posting time and the valence and arousal of their posts in our data set, and further divided by gender. We notice some patterns emerge in our data. Valence increases with age for both genders, especially at the start and end of our age intervals (13–16 and 30–35), confirming the aging positivity bias (Mather and Carstensen, 2005). Valence is higher for females across almost the entire age range. Posts written by females are also significantly higher in arousal for all age groups. Age does not play a significant effect in post arousal, although there is a slight increase with age especially for females. Overall, these figures again illustrate the importance of age and gender as factors to be considered in these types of application (Volkova et al., 2013; Hovy, 2015).

## 3 Predicting Valence and Arousal

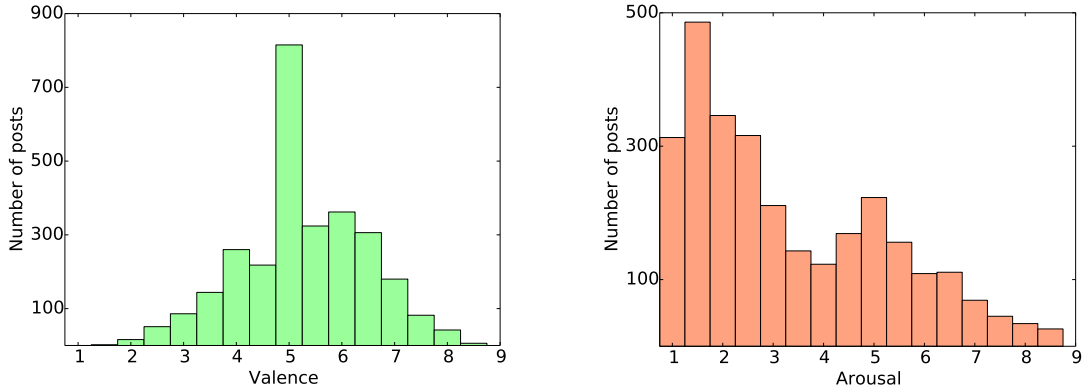
To study the linguistic differences of both dimensions, we build a bag-of-words prediction model of valence and arousal from our corpus.[2] We train two linear regression models with $\ell_2$ regularisation on the posts and test their predictive power in a 10-fold cross-validation setup. Results for predicting the two scores are presented in Table 4.

We compare to a number of different existing general purpose lexicons. First, we use the ANEW (Bradley and Lang, 1999) weighted dictionary to compute a valence and arousal score as the weighted sum of individual word valence and arousal scores. Similarly, we use the affective norms

[2]Available at `http://wwbp.org/data.html`

| Message | V | A |
|---|---|---|
| Is the one whoz GOing to Light Up your Day!!!!!!!!!!!! | 7 | 8 |
| Blessed with a baby boy today ... | 7.5 | 2 |
| the boring life is back :( ... | 3 | 2.5 |
| IS SUPER STRESSED AND ITS JUST THE SECOND MONTH OF SCHOOL ..D: | 2.5 | 7 |

Table 1: Example of posts annotated with average valence (**V**) and arousal (**A**) ratings.




(a) Valence. (b) Arousal.

Figure 1: Histrograms of average rating scores.

of words obtained by extending ANEW with human ratings for ∼14000 words (Warriner et al., 2013). We also benchmark with standard methods for estimating valence from sentiment analysis. First, we use the MPQA lexicon (Wilson et al., 2005), which contains 7629 words rated for positive or negative sentiment, to obtain a score based on the difference between positive and negative words in the post. Second, we use the NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013), which obtained the best performance on the Semeval Twitter Sentiment Analysis tasks.[3]

Our method achieves very high correlations with the target score. Arousal is easier to predict, reaching $r = 0.85$ correlation between predicted and rater score. ANEW obtains significant correlations with both of our ratings, however these are significantly lower than our model. The extended list of affective norms obtains, perhaps surprisingly, lower correlation for valence, but stronger correlation with arousal than ANEW. For valence, both sentiment analysis lexicons provide better performance

[3] https://www.cs.york.ac.uk/semeval-2013/task2/

| Method | Valence | Arousal |
|---|---|---|
| **ANEW** | .307 | .085 |
| **Aff Norms** | .113 | .188 |
| **MPQA** | .385 | – |
| **NRC** | .405 | – |
| **BOW Model** | .650 | .850 |

Table 4: Prediction results for valence and arousal of posts reported in Pearson correlation on 10-fold cross-validation for the BOW model.

than the affective norms lexicons, albeit lower than our model trained on parts of the same data set.

The performance improvement is most likely driven by the domain of the data set. While our method is trained on held-out data from the same domain in a cross-validation setup, the other methods suffer from lack of adaptation to this domain. The NRC lexicon, trained for predicting sentiment on Twitter, obtains the highest performance of the established models, due to the fact that is trained on a more similar domain. The lower performance of the existing models can also be explained by the fact that they predict a score used for *classification* into positive vs. negative, while our target score repre-

sents the *strength* of the positive or negative expression. Moreover, the affective norms scores are handcrafted dictionaries where the weights assigned to words are derived in isolation of context, contain no adaptations to new words, spellings and to the language use from Facebook.

## 4 Qualitative Analysis

In this section we highlight the most important unigram features for each dimension as well as the qualitative difference between the two dimensions of valence and arousal. To this end, we show the words with the highest univariate Pearson correlation with either of the two dimensions in Table 5. Each score is represented by the mean of the two ratings.

| | **Valence** | $r$ | **Arousal** | $r$ |
|---|---|---|---|---|
| Positive | ! | .251 | ! | .773 |
| | :) | .237 | Birthday | .097 |
| | Birthday | .212 | Happy | .081 |
| | Happy | .197 | Its | .079 |
| | Thank | .196 | Wishes | .076 |
| | Great | .195 | Soooo | .074 |
| | Love | .195 | Thanks | .073 |
| | Thanks | .179 | Christmas | .071 |
| | Wishes | .170 | Sunday | .069 |
| | Wonderful | .159 | Yay | .064 |
| Negative | Hate | -.163 | [..]* | -.206 |
| | :( | -.159 | . | -.164 |
| | ? | -.117 | Status | -.064 |
| | Sick | -.112 | Life | -.064 |
| | Why | -.102 | People | -.060 |
| | :'( | -.094 | Bored | -.059 |
| | Not | -.093 | :/ | -.056 |
| | Bored | -.092 | Of | -.056 |
| | Stupid | -.089 | Deal | -.056 |
| | ... | -.087 | Every | -.054 |

Table 5: Words most correlated positively and negatively with the two dimensions.

The results show that both dimensions have similar top features as well as distinct ones. Tokens such as '!', 'Happy', 'Birthday', 'Thanks', 'Wishes' are indicative of both positive valence and arousal, while tokens like 'Bored' and '...' are indicative of both negative valence and low arousal. We notice however tokens that are only indicative of positive valence ('Wonderful', 'Love'), positive arousal ('Sunday', 'Yay'), negative valence ('Why', 'Stupid') or negative arousal ('Life', 'Every', 'People'). The question mark is correlated to negative valence, together with the word 'Why', showing that questions on Facebook are usually negative in valence. Also in terms of punctuation, positive valence and arousal is expressed through exclamation marks, while negative valence and especially arousal is expressed through repeated periods. This behavior is specific to Social Media and which standard emotion lexicons usually does not capture.

Emoticons also exhibit an interesting pattern across the two dimensions. The smiley *:)* is the second most correlated feature with valence, but is not in the top 10 for arousal. Similarly, the frown emoticons (*:(*, *:'(*) are amongst the top 10 features correlated with negative valence, but have no relationship with arousal. The only emoticon correlated highly with low arousal is the undecided emoticon (*:/*).

## 5 Conclusion

In this work, we introduced a new corpus of Social Media posts mapped to the circumplex model of affect. Each post is annotated by two annotators with a background in psychology on two independent nine point scales of valence and arousal, who were calibrated before rating the statuses. We described our annotation process and reviewed the annotation guidelines. In total, we annotated 2895 Facebook posts, discarding the un-ratable ones. The corpus and our valence and arousal bag-of-words prediction models are publicly available.

The results of the annotations have very high agreement. A linear regression model using a bag of words representation trained on this data achieves high correlations with the outcome annotations, especially when predicting arousal. Standard sentiment analysis lexicons predicted both dimensions with lower accuracies.

Our system can be further improved by leveraging the vast amount of available data for Twitter sentiment analysis. We consider this model extremely useful for computational social science research that aims to measure individual user valence and arousal, its relationship to demographic traits and its changes over time or in relation to certain life events.

## Acknowledgements

The authors acknowledge the support of the Templeton Religion Trust, grant TRT-0048.

## References

Margaret Bradley and Peter Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual, and Affective Ratings. Technical report.

Samuel Brody and Noemie Elhadad. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 804–812.

Rafael Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

Maria Gendron and Lisa Feldman Barrett. 2009. Reconstructing the Past: A Century of Ideas about Emotion in Psychology. *Emotion Review*, 1(4):316–339.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 752–762.

Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Greg Park, Lyle H Ungar, David J Stillwell, Michal Kosinski, Lukasz Dziurzynski, and Martin EP Seligman. 2014. From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology*, 50:178–188.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 110(15):5802–5805.

Mara Mather and Laura L Carstensen. 2005. Aging and Motivated Cognition: The Positivity Effect in Attention and Memory. *Trends in Cognitive Sciences*, 9(10):496–502.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, SemEval, pages 321–327.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. *Development and Psychopathology*, 17(3):715–734.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 Task 10: Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval, pages 451–463.

James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval, pages 70–74.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC, pages 1556–1560.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 4 of *LREC*, pages 1083–1086.

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

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1815–1827.

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.

Stephen R Wester, David L Vogel, Page K Pressly, and Martin Heesacker. 2002. Sex Differences in Emotion a Critical Review of the Literature and Implica-

tions for Counseling Psychology. *The Counseling Psychologist*, 30(4):630–652.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 347–354.