Emo2Val: Inferring Valence Scores from fine-grained Emotion Values

Alessandro Bondielli, Lucia C. Passaro and **Alessandro Lenci**
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa (Italy)
alessandro.bondielli@gmail.com
lucia.passaro@for.unipi.it
alessandro.lenci@unipi.it

Abstract

English. This paper studies the relationship between the *valence*, one of the psycholinguistic variables in the Italian version of ANEW (Montefinese et al., 2014), and emotive scores calculated by exploiting distributional methods (Passaro et al., 2015). We show two methods to infer valence from fine grained emotions and discuss their evaluation.

Italiano. *Questo lavoro studia la relazione tra la valenza, una delle variabili psicolinguistiche presenti nella versione italiana di ANEW (Montefinese et al., 2014) e degli score emotivi calcolati distribuzionalmente (Passaro et al., 2015). Mostriamo due metodi per inferire la valenza a partire da tali valori e ne discutiamo la valutazione.*

1 Introduction

Recent years have seen a surge in studies concerning emotional ratings, both in psycholinguistics and in affective computing. Traditionally, the three main behavioral dimensions to measure the emotional value of a word are *valence*, *arousal* and *dominance*. Warriner et al. (2013) define valence as the "pleasantness of the stimulus", usually ranging from 1 (very unpleasant) to 9 (very pleasant). The word *dead* has a low valence rating, whereas *holiday* has a higher one. Arousal is the intensity of the feeling evoked on a scale from "stimulated" to "unaroused". A highly stimulating word is *passion*. On the contrary, *sleep* is not arousing. Finally, dominance is identified with the degree to which the stimulus makes the reader feel "in control" (Louwerse and Recchia, 2014). *Victory* is a word with high dominance.

In the domain of Affective Computing, the goal moves from the identification of such variables to the annotation of the texts with the emotions they express and - for Sentiment Analysis - with their degree of positivity and/or negativity.

The aim of this work is to study the relationship between the most important psycholinguistic variables and emotive scores calculated by exploiting distributional methods. In particular, we will focus on valence ratings, assuming that, within these three dimensions, valence is the most highly related with a positive, negative or neutral emotional content. In fact, it can be defined as the "the polarity of emotional activation" (Lang et al., 1999).

A possible approach to infer the valence of the words from co-occurrence statistics is the one adopted by Louwerse and Recchia (2014), who followed a bootstrapping method to extend the ANEW lexicon (Bradley and Lang, 1999). Another approach would be to exploit a resource such as SenticNet (Cambria et al., 2016) to infer valence based on values of polarity for words or conceptual primitives. An alternative strategy is to infer the valence from an emotive lexicon such as ItEM (Passaro et al., 2015; Passaro and Lenci, 2016), a distributional lexicon for Italian, in which words are associated with an emotive score for 8 different emotions. In our opinion, this solution has several advantages: first of all, ItEM has been proven to be quite robust, and guarantees high coverage over Italian words; secondly, it is not only a static resource, but it can be easily expanded with new words, allowing for a quick adaptation to different contexts. Finally, associating words with fine-grained emotional values allows for a wide range of analyses, such as for instance hate and violence detection in texts.

Experimental results showed, in an indirect way, that distributional emotive ratings can be very useful in the implementation of systems for polarity classification (Passaro and Lenci, 2016;

Bondielli, 2016). However, what is the real relation between emotive scores and valence? Our hypothesis is that emotions can be seen as a representation of valence on a more granular scale. The Plutchik's emotion taxonomy (Plutchik, 1994; Plutchik, 2001) is partitioned into positive or negative emotions. However, borderline emotions such as SURPRISE are harder to be included into a positive or negative class, and therefore to be attributed with a direct valence rating. Words like *party* and *gun* will have widely differing valence ratings, but both strongly elicit the emotion of SURPRISE. Hence it is interesting to ask the following question: given ItEM, are we able to predict the valence (i.e., positivity and/or negativity) of its words? In order to address this latter point, we performed a simple regression model to predict the valence ratings of words in ANEW (Montefinese et al., 2014) given the respective emotive values in ItEM (Passaro et al., 2015; Passaro and Lenci, 2016).

This paper is organized as follow: in Section 2 we describe the resources used for the creation of the model. Section 3 shows our method and the results obtained. Finally, in Section 4 we evaluate the results and discuss our findings.

2 Resources

The main resources we used for our experiments are the Italian version of the Affective Norms for English Words (Montefinese et al., 2014) and the Italian EMotive lexicon (Passaro et al., 2015).

2.1 Italian ANEW

ANEW (Affective Norms for English Words) (Bradley and Lang, 1999) is a database created from a rating of 1034 English words with values for *valence*, *arousal* and *dominance*. Montefinese et al. (2014) provided an Italian version of ANEW, developed by translating the English ANEW words, and by adding the words taken from the Italian semantic norms (Montefinese et al., 2012), for a total of 1121 words. Ratings have been obtained via an experiment where participants had to rate words for the target variables. The reported ratings are the average of the ratings for all participants.

2.2 ItEM

ItEM (Passaro et al., 2015; Passaro and Lenci, 2016) is an emotive lexicon for Italian, in which each target term is associated with a score quantifying its association with each emotion in the Plutchik's taxonomy (Plutchik, 1994): JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE and ANTICIPATION. The resource has been created as follows: in a first phase, feature elicitation was used to create a small set of seed lemmas highly associated to one or more of the emotions in the taxonomy. Then, these lemmas have been distributionally expanded with the most frequent words in two Italian corpora (Baroni et al., 2004; Baroni et al., 2009). Finally, the emotive scores for each word were calculated by measuring the cosine similarity between the lemma and eight emotive centroids built from the collected seeds.

3 From fine-grained Emotion Values to Polarity

We used 2 main regression models to predict the valence from the distributional emotive scores. The first experiment, described in section 3.1 shows a polynomial regression model, and the second one (section 3.2) shows a logistic model in which the valence scores in ANEW have been discretized into two classes representing the positiveness and negativeness of the word.

A simple preprocessing phase has been applied to align the two resources. ANEW has 1121 words, but 65 of them have multiple POS (e.g. *aereo* (plane) can be both a noun and an adjective). We duplicated each word, extending the dataset to 1189 elements, and extracted distinct emotive scores for each <lemma,PoS> pair. In addition, we replaced word forms like "scorie" (waste), with their most frequent word type (scoria) in ItaWaC (Baroni et al., 2004) and La Repubblica (Baroni et al., 2004). Eventually, 57 ANEW words were left out of the analysis because they were not in ItEM. Overall, the resulting size of the aligned dataset is 1129 elements. Finally, to cope with the different distribution of data among the various emotions in ItEM, we normalized the scores with their z-score.

3.1 Polynomial regression

Due to the bimodal distribution of the data in ANEW, we decided to use a polynomial regression model to predict the valence of the words in ANEW by exploiting their emotive normalized scores in ItEM. Preliminary tests had in fact shown that a simple multiple linear regression model was not able to properly fit the data. The histogram

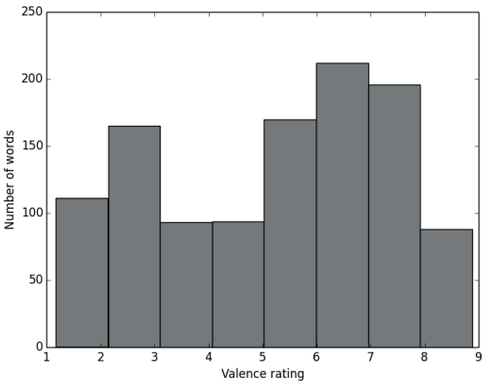

Figure 1: Valence ratings distribution

in Figure 1 shows such data distribution, in which most of the ANEW words have a valence score in the ranges 2-3 and 6-8, with a slight bias towards higher values.

To define the most performing degree (Deg) of the polynomial function, we performed 10-fold cross validation for degrees in the range $\{1...5\}$. The results, presented in Table 1, clearly show overfitting for degrees equal or higher than 3. This is due to the fact that, given the number of parameters (#P), the estimated minimum number of observations (Min. Obs.), computed as $\#P \times 15$, must be at most around the total number of observations. This is true only for polynomial of degree 1 and 2. This finding is in line with Schmidt (1971) and Harrell (2001) who demonstrated that to guarantee the reliability of the prediction, each parameter in the regression model should have a minimum number of observations between 10 and 20.

Deg	#P	Min. Obs.	R^2	MSE
1	9	~ 135	0.46	2.24
2	45	~ 675	0.53	1.82
3	165	~ 2475	0.31	1.50
4	495	~ 7425	-81.29	0.96
5	1287	~ 19305	-11 B	0.00

Table 1: Experiments performed to define the most performing Deg for the polynomial

Given this result, we performed a polynomial interpolation over our parameters with a polynomial of degree 2. Then, we applied a simple multiple linear regression over the new data for predicting the valence. Figure 2 shows the result of the regression fitting. For this model, we obtained a R-Squared (R^2) of 0.58, a mean absolute error (MeanAE) of 1.08, a mean squared error (MSE) of 1.81, and a Median absolute error (MedianAE) of 0.95.

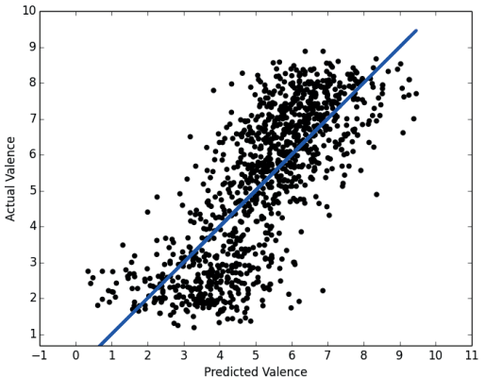

Figure 2: Fitting of predictions

For this experiment, we also provide two additional evaluations (the corresponding results are shown in Table 2):

A) the results of prediction by means of a 10-fold cross validation;

B) the results of prediction by means of split of the data between training (66%) and test (33%).

Method	R^2	MeanAE	MSE	MedianAE
A	0.53	1.13	1.99	0.98
B	0.54	1.13	2.00	0.93

Table 2: Results of the evaluations

We would like to notice that our prediction performs better for words with a very high arousal. In fact, emotionally arousing words were more likely to be produced as an emotive *prototypical* word in the elicitation phase of ItEM. As a consequence, since ItEM's emotive centroids have been constructed using the vectors of these words (namely the seeds), also their nearest neighbors (i.e., the most emotive words) are assumed to have a high level of arousal. Moreover, the distribution of the data in Figure 3, clearly shows how, in ANEW, high arousal corresponds to very high (or very low) valence ratings, suggesting that highly arousing words tend to be very positive or very negative (i.e. polarized). Building on this evidence, we performed an additional experiment in which we used the portion of the data (573 words) with an arousal

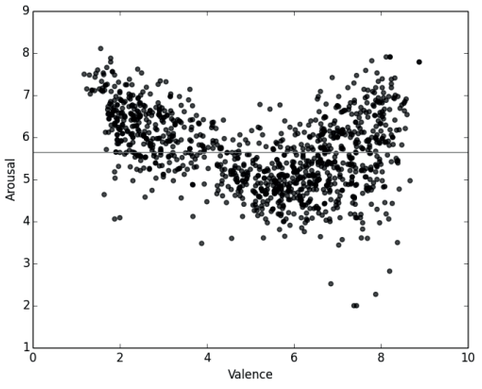

Figure 3: Valence-Arousal distribution

rating higher than its median (5.64) for prediction. In such model, in fact, R^2 is attested to ~ 0.64.

Given the distribution of the data showed in Figure 2, it is clear that a polynomial regression might not be a perfect fit for valence ratings. Nevertheless, it is very important to focus on MeanAE and MSE values. These errors are relatively low with respect to the scale of the human-rated valences.

This means that, on average, the difference between human-rated valence and predicted valence is between 1 and 2. To prove this point, we also compared the obtained scores with the original human annotations, by exploiting the standard deviation for each valence rating. We found that $73,5\%$ of our predictions fall into the correct range around the average valence. If we consider a word having (in ANEW) a valence score of around 8 (e.g. *pace* (peace)) the system will predict a score between 6 and 9, leaving the word around the same (positive) area of the distribution. The same (and opposite) goes for low-valenced words, such as *drogato* (drug addicted) and *feccia* (scum). Problems arise in the case of the words with a medium valence. Examples can be *corridoio* (corridor) and *insipido* (bland). In this case, the word will have the same chance to be attributed with a high valence score (5-6) or with a low one (3-4). Supposing to discretize valence ratings in two classes, a positive and a negative one, with a cut on the median, predictions will fall in the right class for most of the high (or low) valenced words, and (possibly) in the wrong one for the words of medium valence. In fact, by constructing a shallow mapping of the valence into positive (with $valence >= 5.5$) and negative class, we found a correlation of 0.73 between predicted and actual data.

3.2 Logistic regression

Building on the last experiment, and supposing a discretization of the valence into the positive and negative class, we also used a logistic regression model to predict this *binary valence*. The results of this experiment are very promising. We performed 10-fold cross validation to evaluate the effectiveness of the logistic regression over the transformed valence ratings, and obtained an average mean accuracy of 0.80. Detailed results for this evaluation are shown in Table 3.

	Precision	**Recall**	**F1**
MicroAVG	0.806	0.803	0.802
MacroAVG	0.803	0.803	0.803

Table 3: Logistic regression (Cross Validation)

4 Results and discussion

The results provided in previous experiments showed both pros and cons of this approach.

The main advantage of exploiting distributional emotive scores to predict the word's valence is that such scores can be easily obtained in an unsupervised way by means of co-occurrence statistics.

Moreover, predicted data showed a rather good accuracy with respect to the actual distribution, especially considering the logistic regression experiment. In fact, our models reach peak performances by focusing the analysis on the sign of the valence with logistic regression instead of working with continuous values.

On the other hand, the main drawback of our approach derives from the dimension of the ANEW dataset, and in particular from the lack of examples around the medium valence score ratings. It is clear that the ratings distribution in this resource prevented us from obtaining reliable results for continuous values. This might also provide an explanation for the errors concerning the logistic regression experiment. We are confident that having access to a new resource covering the full spectrum of the valence more evenly would have a positive impact on our model.

5 Conclusions and ongoing research

In this work we studied the relationship between *valence* and distributional emotive scores. We modeled our data with regression in order to predict both a continuous score for valence and its corresponding binarized version (i.e., polarity).

Despite the difficulties of modeling an accurate representation of a continuous valence rating from a small and unbalanced dataset like the Italian ANEW, we can identify a clear relationship between distributional emotional scores and a discrete valence obtained by categorizing the ratings into a positive and a negative class.

In the near future, we plan to improve our regression models, with the aim of reducing the impact of the distribution of the data in ANEW, possibly implementing new strategies able to cope with non linear data. ANEW is a highly renown psycholinguistic dataset, but we plan to extend the present work to predict sentiment polarity scores taken from SentiWordNet (Esuli and Sebastiani, 2006a; Esuli and Sebastiani, 2006b), thereby exploiting the larger coverage of this resource.

Moreover, we plan to follow the approach employed in ItEM to create a polarity lexicon for Italian, using ANEW words as seed to build positive and negative polarity centroids. This would also be beneficial for evaluating performances on a emotion-based approach and a polarity-based one.

Finally, we aim at testing the effectiveness of our system for Sentiment Polarity Classification.

References

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the la repubblica corpus: A large, annotated, tei (xml)-compliant corpus of newspaper italian. *issues*, 2:5–163.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Alessandro Bondielli. 2016. Da facebook a twitter: Creazione e utilizzo di una risorsa lessicale emotiva per la sentiment analysis di tweet. Master's thesis, University of Pisa, Italy.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*, pages 2666–2677.

A. Esuli and F. Sebastiani. 2006a. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06)*, Trento (Italy). Association for Computational Linguistics.

A. Esuli and F. Sebastiani. 2006b. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422, Genoa (Italy). European Language Resource Association (ELRA).

F.E. Harrell. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1999. International affective picture system (iaps): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*, 2.

MM Louwerse and G Recchia. 2014. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68(12):1–15.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2012. Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, pages 1–22, oct.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.

Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoro (Slovenia).

Lucia C Passaro, Laura Pollacci, and Alessandro Lenci. 2015. Item: A vector space model to bootstrap an italian emotive lexicon. *CLiC it*, 60(15):215.

Robert Plutchik. 1994. *The psychology and biology of emotion.* HarperCollins College Publishers.

R. Plutchik. 2001. The nature of emotions. *American Scientist*, 89:344–350.

Frank L Schmidt. 1971. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31(3):699–714.

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.