Learning Affect with Distributional Semantic Models

Lucia C. Passaro*
Università di Pisa

Alessandro Bondielli**
Università degli Studi di Firenze

Alessandro Lenci†
Università di Pisa

The affective content of a text depends on the valence and emotion values of its words. At the same time a word distributional properties deeply influence its affective content. For instance a word may become negatively loaded because it tends to co-occur with other negative expressions. Lexical affective values are used as features in sentiment analysis systems and are typically estimated with hand-made resources (e.g. WordNet Affect), which have a limited coverage. In this paper we show how distributional semantic models can effectively be used to bootstrap emotive embeddings for Italian words and then compute affective scores with respect to eight basic emotions. We also show how these emotive scores can be used to learn the positive vs. negative valence of words and model behavioral data.

1. Introduction

In recent years, cognitive science and computational linguistics have seen a rising interest in subjectivity, opinions, feelings and emotions. In psycholinguistics, *valence*, *arousal* and *dominance* are considered the three main dimensions to measure the emotional value of a word. Warriner, Kuperman, and Brysbaert (2013) define these dimensions as follows. Valence is "pleasantness of the stimulus", usually ranging from 1 (very unpleasant) to 9 (very pleasant). An example of a word with low valence is *dead*, whereas *holiday* has an high value. Arousal is instead the *intensity* of the feeling evoked, on a scale from "stimulated" to "unaroused". *Passion* is a highly arousing word, whilst *sleep* is not arousing. Finally, dominance is identified as the degree of "control" felt by a reader given the word as stimulus (Louwerse and Recchia 2014). For example, *victory* is a word with a very high dominance rating. In computational linguistics, the goal moves from the investigation of such psycholinguistic variables at the lexical level to the classification of texts with respect to the emotions they express or, in the case of Sentiment Analysis, to their affective valence.

It is clear that these research areas are closely interrelated, but unfortunately they often tend to ignore each other and to use different methods to create, extend and evaluate their resources. The aim of this work is to show how distributional semantic models can

* Computational Linguistics Laboratory, Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi) - Via S. Maria 36, 56126 Pisa, Italy E-mail: lucia.passaro@fileli.unipi.it
** Dipartimento di Ingegneria dell'Informazione (DINFO) - Via Santa Marta 3, 50139 Firenze, Italy E-mail: alessandro.bondielli@unifi.it
† Computational Linguistics Laboratory, Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi) - Via S. Maria 36, 56126 Pisa, Italy E-mail: alessandro.lenci@unipi.it

be used to bootstrap emotive *embeddings* for Italian words and then compute affective scores with respect to eight basic emotions. We also show how these emotive scores can be used to learn the positive vs. negative valence of words to model behavioral data. We will test the results on human-based ratings, assuming that the rated valence can be defined as the "polarity of emotional activation" (Lang, Bradley, and Cuthbert 1997).

One possible approach to infer valence ratings of 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 is to exploit a resource such as SenticNet (Cambria et al. 2016) to infer valence based on values of polarity for words or conceptual primitives. As shown in Bondielli, Passaro, and Lenci (2017), a third viable strategy is to infer word valence from an emotive resource such as ItEM (Passaro, Pollacci, and Lenci 2015; Passaro and Lenci 2016), a distributional lexicon for Italian, in which words are associated with an emotive score for 8 different emotions. This solution has several advantages. Firstly, ItEM is based on an unsupervised method to estimate affective scores, that guarantees high coverage over Italian words and can be easily expanded, allowing for a quick adaptation to different contexts. Moreover, associating words with fine-grained emotional values allows for a wide range of analyses, such as for instance hate and violence detection in texts.

The vectors used in Bondielli, Passaro, and Lenci (2017) relies on a classical count-based distributional model, and have provided interesting results. Based on these findings, in this work we focus on whether and how results could be improved by exploiting word embeddings learnt with a prediction-based model (Lenci 2018) to compute the affective scores. The proposed strategy is expected to perform better than the count-based one, and it is backed by an extensive body of related work in which Sentiment Lexicons are created by exploiting dense word vector representations obtained with neural network models (Bengio et al. 2003; Mikolov et al. 2013a, 2013b; Turian, Ratinov, and Bengio 2010; Huang et al. 2012). Moreover, such an approach has been successfully implemented in several Sentiment Analysis tasks (Tang et al. 2014; Yu et al. 2017; Castellucci, Croce, and Basili 2015, 2016; Basili, Croce, and Castellucci 2017).

This paper is organized as follows: In section 2 we describe the resources employed in this research, namely ANEW (section 2.1) and ItEM (section 2.2). In section 2.3, we present additional versions of ItEM based on neural embeddings. In section 3, we present three methods to infer valence ratings starting from distributional emotive scores: In the first two experiments, like in Bondielli, Passaro, and Lenci (2017), we predict a continuous valence score by exploiting a polynomial regression model (section 3.1) and a discrete score by means of logistic regression (section 3.2). In a third experiment (section 3.3), we present a new method that uses emotive seeds to predict a word valence. In this latter case, to assess the reliability of the method, we measure the correlation between the predicted scores and the human rated ones in ANEW. All experiments have been carried out with the count-based and the prediction-based versions of ItEM, to compare the effect of these two families of distributional models to learn the affect of lexical items. Finally, in section 4 we discuss our results and findings.

2. Affective resources

The main goal of this paper is to show that distributional emotive and affective scores can be used to infer a word's valence, as a crucial piece of information to determine the affective content of texts. Our research relies on two main resources, which we describe in this section: The Italian version of the Affective Norms for English Words (ANEW)

(Montefinese et al. 2014) and the Italian EMotive lexicon (ItEM) (Passaro, Pollacci, and Lenci 2015; Passaro and Lenci 2016).

2.1 Italian ANEW

ANEW (Bradley and Lang 1999) is a database containing 1034 English words rated for *valence*, *arousal* and *dominance*. The affective rating system used to annotate words is a variant of the Self-Assessment Manikin (SAM: Lang (1980)). The SAM is a technique built with the aim to assess the affective reaction of a person to different kinds of stimuli, in terms of pleasure, arousal and dominance (Bradley and Lang 1994). In ANEW, the SAM uses a numerical scale, ranging from 1 to 9, and is applied to all the main variables. For example, if we consider the valence, the rate 1 means unpleasant and 9 means very pleasant.

Connotation is a cultural phenomenon that may vary greatly between languages and different time spans (Das and Bandyopadhyay 2010) and, consequently, the "correct" translation of a word can have a different emotional connotation in different languages (Chen, Kennedy, and Zhou 2012). For this reason, the collection of affective norms has been carried out for many languages including Italian. The Italian adaptation of ANEW contains the norms for the translation of the original ANEW words, as well as for words taken from the Italian Semantic Norms (Montefinese et al. 2013). The total number of annotated words is 1,121. The three main dimensions of valence, arousal and dominance were rated using again the SAM scale, in order to provide consistency with the original norms. Apart from the original affective ratings, new dimensions were collected as well, namely subjective and objective psycholinguistic indexes. Subjective indexes are *familarity*, *imageability*, and *concreteness*. The familiarity index is based on subjective measures of how often participants both use and are exposed to a given word (Montefinese et al. 2014); Concreteness is the extent to which a word is tangible (Paivio, Yuille, and Madigan 1968); Imageability refers to the ease of generating a mental image for a word (Paivio, Yuille, and Madigan 1968). Objective indexes represent features of a word, such as length, frequency in two corpora (CoLFIS (Bertinetto et al. 2005) and La Repubblica (Baroni et al. 2004)), and number of orthographic neighbors. For the affective ratings, researchers also held into account gender differences. For example, a word like *allegro* "merry" has a very high rating for valence (8.11), and relatively high ratings for arousal and dominance (5.89 and 6.86 respectively). On the other end of the spectrum, *afflizione* "grief" is rated very low for valence (1.94), but it is considered a medium-high arousing word (6.39) and a medium-low dominance word (3.18).

The experiments were conducted on 1,084 participants, all native speakers and undergraduate psychology students. Out of all the participants, 684 were used to rate words with valence, arousal and dominance scores, and 400 to perform familiarity, imageability and concreteness evaluations. Each word was rated by at least 31 participants (of whom at least 10 male) for affective ratings, and by at least 20 participants for psycholinguistic ratings. Participants were asked to rate words using the SAM scale for affective ratings. The final resource is therefore composed of the original ANEW word, its Italian translation, and mean scores and standard deviation for each of the considered dimension. For affective ratings, measurement are also reported for male and female participants.

The main contribution of the Italian ANEW to the present research is that it provides us with an highly controlled scoring for affective ratings, that can be easily exploited to evaluate affective distributional scores.

2.2 ItEM

The Italian EMotive lexicon (ItEM) is a distributional resource described in Passaro, Pollacci, and Lenci (2015), Passaro and Lenci (2016), and based on the so-called Distributional Hypothesis (Harris 1954), which states that semantically similar words tend to appear in similar contexts. In ItEM, this hypothesis has been generalized to emotions, as follows:

A word w is associated with an emotion e if it co-occurs in similar contexts of other words associated with e.

To implement this hypothesis, in the basic version of ItEM, each emotion has been represented as a centroid vector built out of a set of seed words strongly associated to each of the target emotions.

The resource has been developed in a three stage process. The first phase was devoted to the collection a small set of seed words highly associated with one of Plutchik's basic emotions (Plutchik 1980): JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE and ANTICIPATION. In a second phase, distributional semantic methods were exploited to expand the seeds and populate the resource. Finally, the automatically extracted emotive annotations have been evaluated via crowdsourcing.

The goal of the first phase was to collect a small lexicon of *emotive lexemes*, highly associated to one or more Plutchik's basic emotions. Such a goal was reached by means of an online feature elicitation paradigm, in which 60 Italian native speakers were asked to list, for each emotion, 5 lexical items for each of our PoS of interest (Nouns, Adjectives and Verbs). After applying various filters and revisions, we obtained a lexicon of 347 words. For each word in this set, its emotion distinctiveness score was calculated – following Devlin et al. (1998) – as its informativeness (i.e., the reciprocal of the number of emotions for which the word was generated). For example, the distinctiveness of the word *amore* "love" is 1/3, given the following distribution of its production frequency: JOY = 2, TRUST = 5, and ANTICIPATION = 4. The seeds were restricted to the words with a distinctiveness score equal to 1 (i.e., the words produced/evoked by a single emotion). In addition, this set was expanded with the names of the emotions such as *gioia* "joy" , *rabbia* "anger" and their synonyms attested in Multiwordnet (Pianta, Bentivogli, and Girardi 2002) and Treccani Online Dictionary[1] for a total of 555 emotive seeds.

In the bootstrapping phase, a count-based Distributional Semantic Model (DSM) was used to expand the seeds using a corpus-based model inspired to Turney and Littmann (Turney and Littman 2003) to automatically infer the semantic orientation of a word from its distributional similarity with a set of positive and negative words. In particular, the DSM was created by extracting from La Repubblica corpus (Baroni et al. 2004) and itWaC (Baroni et al. 2009) the list of the 30,000 most frequent nouns, verbs and adjectives and recording their co-occurrences within a five word symmetric window centered on the target word. Co-occurrences were reweighted with Positive Pointwise Mutual Information (PPMI) (Church and Hanks 1990), but with negative values raised to 0. To optimize the vector space, we followed the approach in Polajnar and Clark (2014) and we selected the top 240 contexts for each target word. As a last step, we applied singular value decomposition (SVD), to reduce the matrix to 300 dimensions.

Adapting the approach Turney and Littman (2003), the emotions were represented as centroid vectors built from the mean of the vectors of the relative seeds. For each

1 http://www.treccani.it/vocabolario/.

emotion E, we computed a word emotive score σ by measuring the cosine similarity of the word vector $\overrightarrow{\boldsymbol{w}}$ in the DSM with the centroid vector of E ($\overrightarrow{\boldsymbol{C}_{\mathrm{E}}}$):

$$\sigma(E, w) = \frac{\overrightarrow{\boldsymbol{w}} \cdot \overrightarrow{\boldsymbol{C}_{\mathrm{E}}}}{||\overrightarrow{\boldsymbol{w}}|| \cdot ||\overrightarrow{\boldsymbol{C}_{\mathrm{E}}}||} \tag{1}$$

This score measures the association of a word with a given emotion. For instance, the amount of ANGER associated with the noun *gelosia* "jealousy" is estimated with the cosine similarity between the vector of *gelosia* and the centroid vector of ANGER. The following is the emotion distribution of that word, modeled with the cosine similarity with the emotive centroids: ANGER: 0.65; DISGUST: 0.43; FEAR: 0.36; SADNESS: 0.32; JOY: 0.24; SURPRISE: 0.24; ANTICIPATION : 0.20; TRUST : 0.12.

ItEM was evaluated with two crowdsourcing tasks on the Crowdflower (CF) platform[2] to compare the model performance on a random set of words, including also possibly neutral words, associated with human ratings about their association or lack of association with emotions. The details and results of the ItEM evaluation are reported in Passaro and Lenci (2016).

2.3 Adapting ItEM to prediction-based word embeddings

In order to adapt ItEM to prediction-based word embeddings, we developed a new model, namely the ITEM-8-PREDICT, in which the vectors of the words were built with the state-of-the-art prediction-based DSM Word2vec (Mikolov et al. 2013a, 2013b). In particular, the neural word embeddings were trained on the lemmatized concatenation of the corpora La Repubblica (Baroni et al. 2004) and itWaC (Baroni et al. 2009), by restricting the vocabulary to nouns verbs and adjectives and representing each token in the form $< lemma - PoS >$. After testing few configurations, we used the Skip-Gram with Negative Sampling algorithm with the following hyperparameters: the size of the embedding was set to 500 for each word; the context span was set to 5; the occurrence threshold was set to $1 * e^{-4}$, and the number of negative examples was set to 10. For the sake of comparison, we decided to implement a 500 dimensions vector for the count model as well, which will be referred as ITEM-8-COUNT for the rest of the paper.

3. From fine-grained Emotion Values to Polarity

To predict valence ratings from the distributional emotive scores, we performed several experiments. In Bondielli, Passaro, and Lenci (2017), we showed two alternative methods to predict, respectively, a continuous and a discrete valence rating by exploiting distributional emotive scores. In particular, we used a polynomial and a logistic regression model to infer valence from emotions. In this work we explore this problem more deeply, and propose new distributional methods to construct valence lexicons.

For the sake of comparison, we conducted our experiments on the same dataset analyzed in Bondielli, Passaro, and Lenci (2017). First of all, a simple preprocessing phase was applied to align Italian ANEW and ItEM. The former includes 1,121 words, but 65 of them have multiple PoS (e.g., *aereo* "plane" can be both a noun and an adjective). We

2 http://www.crowdflower.com.

duplicated each word, extending the dataset to 1,189 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 ItWaC and La Repubblica.

Eventually, in all the experiments, some ANEW words were left out of the analysis because they were not covered by the current version of ItEM. This happened for two reasons. Firstly, the word was not included among the ItEM target terms (i.e., low-frequency words not appearing in the list of the top 30,000 words of the considered corpora). Secondly, the words had a negative cosine values with all the emotive centroids in ItEM. In each experiment, we report the coverage with respect to Italian ANEW.

3.1 Predicting a continuous valence score with polynomial regression

As shown in Figure 1, the distribution of Italian ANEW data is bimodal and therefore, we used a polynomial regression model to predict the valence of words in ANEW with their emotive scores in ItEM.

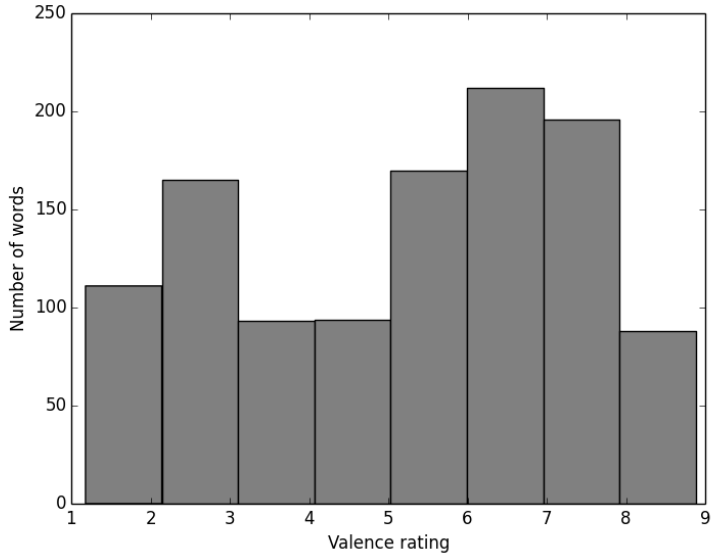

Figure 1
Distribution of valence ratings for Italian ANEW Italian. The histogram clearly shows a lower number of examples for valence ratings in range $[3, 5]$ and for very high and very low values. On the contrary, words with a valence rating in the ranges $[5, 8]$ and $[2, 3]$ are well represented, with a slight bias towards higher values.

To define the most performing degree (Deg) of the polynomial function, we carried out 10-fold cross validation for degrees in the range $[1, 5]$. We can clearly identify overfitting starting from degrees equal or higher than 3 (cf. Table 1). This is due to the fact that, given the number of parameters ($\#P$) for regression, we can estimate the minimum number of observations (Min. Obs.) needed to avoid overfitting. This number was computed as $\#P \times 15$, and should be smaller or equal to the total number of observations used to build the model. In our case, this was true only for polynomial of degree 1 and 2. This finding is in line with Schmidt (1971) and Harrell (2001). In their

work, they demonstrated that to guarantee the reliability of the prediction, for each parameter in the regression model there should be a minimum number of observations between 10 an 20 in the data.

Table 1
Experiments performed to define the best degree (Deg) for the polynomial. For polynomial of degree 1 and 2 we can see an increase in the computed R^2. For higher degrees, the minimum number of observation exceeds the size of our dataset. This causes the MSE to decrease, but R^2 drastically drops as well. The best performing degree for our dataset with respect to R^2 and MSE is degree 2.

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

Given these results, we decided to use a degree 2 for the interpolation of our parameters. We built two models and compared their results. The first model, which we called COUNT, replicates the model presented in Bondielli, Passaro, and Lenci (2017), and exploits the emotive scores in ITEM-8-COUNT, the only difference being vector dimensionality, which was now set to 500. The second model, which we called PREDICT, was built by exploiting the emotive scores in ITEM-8-PREDICT.

We performed polynomial interpolation of the parameters (i.e., the distributional emotive values), and applied a simple multiple linear regression over the new data in order to predict valence. Results of this experiment are shown in Table 2. First of all, we show how the models predict the actual ANEW valence ratings by exploiting the whole dataset. Then, we perform 10-fold cross validation in order to better assess the predictive capabilities of our DSMs. The results, where R-squared (R^2), mean absolute error (MeanAE), mean squared error (MSE), and median absolute error (MedianAE) were used for evaluation, are shown in Table 3.

Table 2
Results of the evaluations. Both models are based on the analysis of 1,090 data points, i.e. the words contained in both ItEM and ANEW. Prediction-based word embeddings show improvements for predicting the whole dataset and for 10-fold cross validation (CV). More specifically, R^2 is increased by 5 points, and all the mean and median errors are reduced.

Model	*R^2*	*MeanAE*	*MSE*	*MedianAE*
COUNT	0.64	0.98	1.54	0.81
COUNT - CV	0.61	1.01	1.65	1.01
PREDICT	0.69	0.89	1.29	0.72
PREDICT - CV	0.66	0.93	1.41	0.93

The results show the same trend for both the COUNT and PREDICT model. We see that the difference between human-rated valence and predicted valence is on average around 1 (it falls between 0.9 and 1.5). However, the results also show that the PREDICT model clearly outperforms the COUNT model for what concerns R^2. This means that a

distributional emotive space such as ItEM can benefit from using prediction-based word embeddings to compute the emotive connotation of words. In addition, the Pearson's correlation coefficient between predicted and human-rated valences is increased from 0.8 ($p < 0.005$) of the COUNT model to 0.83 $p < 0.005$ of the PREDICT one. In both cases, correlation is very high, proving the excellent ability of DSMs to model behavioral data about word affective valence.

It is also important to stress that the use of word embeddings improved the results for MeanAE and MSE. This means that the predictions of the PREDICT model are on average closer to the actual valence ratings of words. This is crucial in order to improve performances for words with medium valence ratings. In fact, the model presented in (Bondielli, Passaro, and Lenci 2017) performed better on low-valenced or high-valenced words. Medium-valenced words on the contrary had more chances to be predicted as either too high or too low, given the mean errors of the model. The PREDICT model, albeit not perfect, may be less prone to this kind of problem, given a generally smaller average error.

3.2 Predicting a discrete valence score with Logistic regression

Following the approach presented in Bondielli, Passaro, and Lenci (2017), we performed a second experiment to evaluate the results of a logistic regression classifier aimed at predicting a discrete valence score. The discretization of the *gold* valence was performed by considering as POSITIVE the words with $valence >= 5.5$, and as NEGATIVE the others. Again, we compared the two versions of ItEM, that is ITEM-8-COUNT and ITEM-8-PREDICT. The goal was to predict a *binary valence* and assess the differences between count- and prediction-based DSMs. Results of these experiments are shown in Table 3.

Table 3
Logistic regression (Cross Validation). Both models are based on the analysis of 1,090 data points, i.e. the words contained in both ItEM and ANEW. Precision, Recall and F1 are improved by exploiting prediction-based embeddings to build ItEM

Model	*Precision*	*Recall*	*F1*
COUNT - MACROAVG	0.828	0.820	0.821
PREDICT - MACROAVG	0.844	0.841	0.842

The results of this experiment again show how the use of neural embeddings can improve the classification performances. The COUNT model has an average F1 of 0.82, whereas the PREDICT model scores 0.84 on the same data.

3.3 Predicting a polarity score with a valence version of ItEM

In a third experiment, we created a distributional polarity lexicon in which Italian words were associated with a positiveness and a negativeness score, rather than the 8 emotive scores described in section 2.2.

First of all, we splitted the emotions into a positive and a negative group. In particular, the seeds elicitated for the emotions JOY and TRUST have been grouped into the class POSITIVE and the seeds elicitated for SADNESS, ANGER, FEAR and DISGUST have been classified as NEGATIVE. The emotions SURPRISE and ANTICIPATION have

been left out of the analysis because of their mixed nature. More specifically we selected the words with a cue validity higher than 0.6 for the original emotion and a production frequency higher than 1. Globally, we selected 119 positive seeds and 310 negative ones. The bootstrapping phase was performed as with the ITEM-8 models. This way, we built two new models, namely ITEM-2-COUNT trained with a count-based DSM (cf. section 2.2), and ITEM-2-PREDICT trained with Word2vec (cf. section 2.3).

To evaluate this method, we approximated the polarity of a word w with the difference between its positiveness (eq. 3) and negativeness (eq. 4) score:

$$polarity(w) = positiveness(w) - negativeness(w) \tag{2}$$

Both scores were calculated as the cosine similarity between the vector of the word ($\overrightarrow{\boldsymbol{w}}$) and the centroid vector of positiveness ($\overrightarrow{\boldsymbol{C}_{\mathrm{P}}}$) and negativeness ($\overrightarrow{\boldsymbol{C}_{\mathrm{N}}}$):

$$positiveness(w) = \frac{\overrightarrow{\boldsymbol{w}} \cdot \overrightarrow{\boldsymbol{C}_{\mathrm{P}}}}{||\overrightarrow{\boldsymbol{w}}|| \cdot ||\overrightarrow{\boldsymbol{C}_{\mathrm{P}}}||} \tag{3}$$

$$negativeness(w) = \frac{\overrightarrow{\boldsymbol{w}} \cdot \overrightarrow{\boldsymbol{C}_{\mathrm{N}}}}{||\overrightarrow{\boldsymbol{w}}|| \cdot ||\overrightarrow{\boldsymbol{C}_{\mathrm{N}}}||} \tag{4}$$

A polarity score close to 1 indicates positiveness while a score close to -1 means negativeness.

In these experiments, we measured the correlation coefficient between Valence and Polarity. Table 4 shows the results of the correlation between the valence in ANEW and the polarity calculated using count-based vs. prediction-based semantic vectors.

Table 4
Correlation coefficient between the Valence in ANEW and the Polarity produced using ITEM-2-COUNT and ITEM-2-PREDICT. We provide both Pearson and Spearman correlation coefficients. In all the experiments we found a $p - value < 0.001$. Both models are based on the analysis of 1,090 data points, i.e. the words contained in both ItEM and ANEW.

Model	*Pearson r*	*Spearman ρ*
COUNT	0.743	0.777
PREDICT	0.785	0.794

The results of this experiment show that the distributional polarity highly correlates with human-elicited data. Moreover, once again the use of word embeddings improves the prediction. These results, compared with the ones obtained with the polynomial regression model (see section 3.1), prove that this method is a reliable alternative to predict valence from polarity, but, at the same time, that a more granular emotion taxonomy, when available, is the best option.

Moreover, by discretizing both valence and polarity with the thresholds used in section 3.2, we observe that the binary models, especially count ones, despite achieving

acceptable accuracy, are outperformed by distributional models relying on a richer emotion taxonomy. The results of this experiment are shown in Table 5.

Table 5
Performance of the discretized model. The discretization of the *gold* valence was performed by considering as POSITIVE the words with $valence >= 5.5$, and as NEGATIVE the others. The polarity was discretized by considering its sign (i.e., the words with a Polarity higher than 0.0 were considered as POSITIVE).

Model	*Precision*	*Recall*	*F1*
COUNT - MACROAVG	0.796	0.791	0.785
PREDICT - MACROAVG	0.827	0.828	0.826

4. Discussion

Overall, our experiments demonstrated that the valence of words can be inferred by means of both emotions and polarity estimated with Distributional Semantic Models. In particular, we showed three methods to infer valence ratings starting from distributional emotive scores: in the first two experiments, inspired by Bondielli, Passaro, and Lenci (2017), we predicted a continuous valence score by exploiting a polynomial regression model (section 3.1) and a discrete score by means of logistic regression (section 3.2). In a third experiment (section 3.3), we showed a method to infer valence directly by exploiting emotive seeds.

All experiments have been carried out with the count-based and the prediction-based versions of ItEM, to compare the effect of these two families of distributional models. In the first experiment we found that the use of word embeddings improves the performance despite the presence of medium valence words, which are supposedly the most difficult to classify. In the second one, we showed that by discretizing the valence into two polarity classes, such an improvement becomes more pronounced, by reaching an F1 of 0.84. Also in this case the model benefits from the use of prediction-based word vectors. Finally, in the third experiment we directly exploited a binary categorization of emotions to infer the valence and we found a high correlation between predicted and human rated valence. However, the model produced by this experiment, albeit being able to reliably predict valence from polarity, suffers from worse performances with respect to the model presented in the first experiment. This demonstrates that a more granular emotion taxonomy might be a better option. Due to the superiority of the PREDICT model in all the experiments we performed, we decided to deeper investigate the differences in terms of correlation between the two distributional methods aimed at predicting a continuous value of valence (experiment 1 and experiment 3). In particular, we studied the effect of the frequency and of the part of speech on the performances of the two models.

For what concerns frequency, we divided the dataset into three equally sized frequency classes. In this case, all the experiments showed the absence of a statistically significant difference between the COUNT and the PREDICT model for low frequency words. In all the other cases, the PREDICT model seems to work significantly better than the COUNT one.

As for Part of Speech, the results are shown in detail in Table 6. Although overall the difference between the PREDICT and the COUNT model is statistically significant

in terms of correlation between the ANEW valence and the distributional polarity, we found such difference to be not significant in the case of verbs.

Table 6
Performance in terms of Pearson's correlation divided by Part of Speech. The table reports the model and the PoS (together with the number of items in the test set).

Model	*Overall [1090]*	*Nouns [782]*	*Verbs [51]*	*Adj [254]*
COUNT (8 EMOTIONS)	0,80	0,79	0,78	0,85
PREDICT (8 EMOTIONS)	0,83	0,83	0,75	0,87
COUNT (2 EMOTIONS)	0,74	0,75	0,71	0,79
PREDICT (2 EMOTIONS)	0,79	0,78	0,71	0,82

Moreover, looking at verbs, we noticed a more pronounced drop in the correlation between actual and predicted values for the PREDICT model with respect to the COUNT one. In other words, the Δ between the correlation of verbs and overall results varies in the range 0.02 and 0.03 for the COUNT model while varying from 0.08 and 0.09 in the PREDICT one. It is clear that the dimension of the sample of the verbs affects the results (especially in the ITEM-8 experiment, in which the points in ANEW are directly embedded in the regression model), but these results open new questions about the behaviour of the prediction-based vectors to model the affective dimension of verbs. This suggests the existence of interesting differences between the two families of DSMs with respect to different PoS, a point we leave for future investigations.

5. Conclusions and ongoing research

In this work we studied the relationship between *valence* and distributional emotive scores inferred from count-based and prediction-based dense semantic vectors. We modeled our data with regression and correlation in order to predict both a continuous score for valence and its corresponding binomial version (i.e., polarity). The results we obtained in our experiments show both pros and cons of each approach. The exploitation of distributional emotive scores for predicting the valence rating for a word may prove advantageous because such scores can be easily obtained in an unsupervised way. Our experiments have in fact shown that, despite using relatively simple models such as polynomial and logistic regression and the creation of a polarity lexicon, we are able to infer valence ratings with good accuracy.

The experiments support two important conclusions:

- prediction-based DSMs produce significantly better lexical representations than count-based ones. This fact was already shown in number of semantic tasks by Baroni, Dinu, and Kruszewski (2014) and Mandera, Keuleers, and Brysbaert (2017). Our research is the first one to prove that this is true also to estimate the affective content of lexical items. Neural embeddings provide on average 3 points of improvement if we consider the Pearson's correlation and of 3 percentage points if we consider the F1 in the prediction of discrete valence;
- most research on Affective Computing focuses on valence defined as a binary category, but it is preferable to rely on a more granular emotion taxonomy, such as the one used by ItEM. Word valence can be better predicted by DSMs trained on 8 basic emotions, rather than DSMs directly trained on seeds grouped into a

positive and a negative class. Of course, this may also depend on the grouping criteria and on the fact that the seeds were originally collected with respect to their association with emotions rather than for their valence. We leave this point to further research.

One of the main drawbacks of our evaluation 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. We are still 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. 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 the seeds used to build our distributional resources and to extend this work to predict sentiment polarity scores taken from SentiWordNet (Esuli and Sebastiani 2006a, 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. In this case, we intend to evaluate the new resource with crowdsourcing or controlled psycholinguistic experiments. Finally, we aim at testing the effectiveness of our system for Sentiment Polarity Classification of texts.

References

Baroni, Marco, 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. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1171–1174, Lisbon, Portugal. European Language Resource Association (ELRA).

Baroni, Marco, 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.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, volume 1, pages 238–247, Baltimore, Maryland.

Basili, Roberto, Danilo Croce, and Giuseppe Castellucci. 2017. Dynamic polarity lexicon acquisition for advanced social media analytics. *International Journal of Engineering Business Management*, 9:1–18.

Bengio, Yoshua, Régean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bertinetto, Pier Marco, Cristina Burani, Alessandro Laudanna, Lucia Marconi, Daniela Ratti, Claudia Rolando, and Anna Maria Thornton. 2005. Corpus e lessico di frequenza dell'italiano scritto (CoLFIS). Technical report.

Bondielli, Alessandro, Lucia C. Passaro, and Alessandro Lenci. 2017. Emo2val: Inferring valence scores from fine-grained emotion values. In *Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 48–52, Rome, Italy. Accademia University Press.

Bradley, Margaret M. and Peter J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Bradley, Margaret M. 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.

Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn W. Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of the 26th*

International Conference on Computational Linguistics: Technical Papers (COLING 2016), pages 2666–2677, Osaka, Japan.
Castellucci, Giuseppe, Danilo Croce, and Roberto Basili. 2015. Acquiring a large scale polarity lexicon through unsupervised distributional methods. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane, and Elisabeth Métais, editors, *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)*, pages 73–86, Passau, Germany. Springer International Publishing.
Castellucci, Giuseppe, Danilo Croce, and Roberto Basili. 2016. A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 38–45, Portorož, Slovenia.
Chen, Stephen H., Morgan Kennedy, and Qing Zhou. 2012. Parents' expression and discussion of emotion in the multilingual family: Does language matter? *Perspectives on Psychological Science*, 7(4):365–383.
Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.
Das, Amitava and Sivaji Bandyopadhyay. 2010. Towards the global SentiWordNet. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 2010)*, pages 799–808, Sendai, Japan.
Devlin, Joseph T., Laura M. Gonnerman, Elaine S. Andersen, and Mark S. Seidenberg. 1998. Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of cognitive Neuroscience*, 10(1):77–94.
Esuli, Andrea and Fabrizio 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 (EACL'06)*, pages 193–200, Trento, Italy. Association for Computational Linguistics.
Esuli, Andrea and Fabrizio Sebastiani. 2006b. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, Genoa, Italy. European Language Resource Association (ELRA).
Harrell, Frank E. 2001. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York.
Harris, Zelig. 1954. Distributional structure. *Word*, 10(23):146–162.
Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers (ACL 2012)*, volume 1, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.
Lang, Peter J. 1980. Behavioral treatment and bio-behavioral assessment: Computer applications. In Joseph B. Sidowski, James H. Johnson, and Thomas A. Williams, editors, *Technology in Mental Health Care Delivery Systems*. Ablex Pub. Corp., Norwood, NJ, pages 119–137.
Lang, Peter J., Margaret M. Bradley, and Bruce N. Cuthbert. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58.
Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.
Louwerse, Max M. and Gabriel 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.
Mandera, Paweł, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Computing Research Repository (CoRR)*.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)*, volume 2, pages 3111–3119, Lake Tahoe, Nevada. Curran Associates Inc.
Montefinese, Maria, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2013. Semantic memory: A feature-based analysis and new norms for italian. *Behavior Research Methods*, 45(2):440–461.

Montefinese, Maria, 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.

Paivio, Allan, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1–25.

Passaro, Lucia C. 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)*, Portorož, Slovenia.

Passaro, Lucia C., Laura Pollacci, and Alessandro Lenci. 2015. Item: A vector space model to bootstrap an italian emotive lexicon. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 215–220, Trento, Italy. Academia University Press.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet (GWC2002)*, pages 293–302, Mysore, India.

Plutchik, Robert. 1980. General psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, 1:3–33.

Polajnar, Tamara and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2014)*, pages 230–238, Gothenburg, Sweden. Association for Computational Linguistics.

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

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, volume 1: Long Papers, pages 1555–1565, Baltimore, Maryland, June. Association for Computational Linguistics.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Management Information Systems (TMIS)*, 21(4):315–346.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Yu, Liang-Chih, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.