Yuqiao Gu, Fabio Celli, Josef Steinberger, Andrew James Anderson,
Massimo Poesio, Carlo Strapparava and Brian Murphy

Using Brain Data for Sentiment Analysis

Abstract

We present the results of exploratory experiments using lexical valence extracted from brain using electroencephalography (EEG) for sentiment analysis. We selected 78 English words (36 for training and 42 for testing), presented as stimuli to 3 English native speakers. EEG signals were recorded from the subjects while they performed a mental imaging task for each word stimulus. Wavelet decomposition was employed to extract EEG features from the time-frequency domain. The extracted features were used as inputs to a sparse multinomial logistic regression (SMLR) classifier for valence classification, after univariate ANOVA feature selection. After mapping EEG signals to sentiment valences, we exploited the lexical polarity extracted from brain data for the prediction of the valence of 12 sentences taken from the SemEval-2007 shared task, and compared it against existing lexical resources.

1 Introduction and related work

Sentiment analysis—automatically recognizing the emotions conveyed by a text, and in particular distinguishing positive from negative valence—has become one of the most popular research areas in computational linguistics (Pang & Lee, 2008; Liu, 2012) both because of the interest of the field in the interplay between emotion and cognitive abilities, and because of its obvious applications (e.g., companies could analyze social networks to determine customer response to their products). Such research however requires collecting judgments about the valence of sentences and possibly lexical items, and simply asking subjects often results in low inter-annotator agreement levels (Arnstein & Poesio 2008; Craggs & McGee Wood, 2004; Esuli & Sebastiani 2006). But this difference between subjective judgments may be caused by strategic effects rather than unconscious processes as measured with neuroimaging techniques. And indeed, Crosson et al. (1999, 2002) and Cato et al. (2004) demonstrated that it is possible to discriminate positive and negative words from neutral words on the basis of the blood-oxygen-level dependent (BOLD) signal collected through functional magnetic resonance imaging (fMRI) scans. Using magnetoencephalography (MEG) recording techniques**,** Hirata et al., 2007 found that negative and positive words can be distinguished by event-related desynchronizations (ERDs). These results suggest that valence information might be best collected without asking the subjects directly. In the future it may be possible to use neuroimaging to benefit sentiment analysis e.g. by tapping into subconscious valence representations which could reduce annotator rating time; or provide us more nuanced ways to measure valence. The long-term aim of our project is to assess the feasibility of using for sentiment analysis valence information derived from the brain.

The focus of the preliminary investigation discussed in this paper was primarily practical: to address one of the issues that have to be faced in order to achieve the ultimate goal. The problem is that the cost of collecting valence information through fMRI or MEG would be prohibitive at present. On the other hand, EEG is a very inexpensive and widespread technology. Taking advantage of its high temporal resolution, in recent years EEG and event-

related potentials (ERPs) was intensively used in psycholinguistics, e.g., for the investigation of processing mechanisms of semantic categories (Pulvermüller *et al*., 1999; Kiefer 2001; Paz-Caballero *et al*., 2006; Proverbio *et al*., 2007; Hoenig, *et al*., 2008; Adorni & Proverbio, 2009; Fuggetta, *et al*., 2009; Renoult & Debruille, 2010; Renoult et al., 2012). Hagoort *et al*. (2004) studied the integration of word meaning and world knowledge with EEG, ERP and fMRI while subjects read sentences. In some sentences the critical words make the sentences a correct or false semantic interpretation and in other sentences the critical words make the sentence a correct or false world knowledge interpretation. Using EEG and ERP, Delong *et al*. (2005) found that individuals can use linguistic input to pre-activate representations of upcoming words in advance of their appearance. Using event-related EEG and multivariate pattern analysis, Simanova *et al*., 2010 studied the conceptual representation and classification of object categories in different modalities. In other work, we have used EEG and machine learning to decode the semantic categories of animals vs tools in younger and elderly subjects during a covert image naming task (Murphy *et al*., 2011; Gu *et al*., 2013). In this work, we apply this approach to the decoding of the emotional valence of written words, and propose a novel paradigm for using such decoding techniques for sentiment analysis.

The structure of the paper is as follows. First of all we describe the paradigm in general terms. Next we discuss how we used a linguistically controlled data set of word stimuli to elicit EEG data about valence and to train a within-subjects valence classifier which was then used to assign valence to words in the test set. Finally, we discuss preliminary experiments using this valence for sentiment analysis.

2 Methodology

A number of issues need to be tackled in order to use brain data to determine the valence of words. The first problem, already mentioned, is that fMRI as used by Cato *et al* is very expensive (the costs are in the order of €500 per hour) and requires substantial medical infrastructure. As already mentioned, our solution to this problem was to use EEG, which costs substantially less and is becoming a standard facility also in Computer Science and Psychology labs.

But even using EEG, it is not possible to get the valence of each word directly from subjects. Generally at least 5-6 presentations of a stimulus (word) to each subject are needed to get a stable representation of the signal for that stimulus and that subject. At a few seconds per stimulus, at most 80 stimuli can be presented to a subject in one hour—the duration of time after which the subject's attention generally is lost. This makes it time-consuming to measure brain activity for even the relatively small number of words in a standard corpus. Creating an EEG-based sentiment dictionary would require multiple sessions for multiple participants. In these experiments we used a test subset of the corpus created for the Sentiment Analysis at SemEval-2007 (Strapparava & Mihalcea, 2007) as test data. The corpus consists of about 250 examples of news titles in the trial set and about 1000 in the test set. News titles have been extracted from news web sites (such as Google news, CNN) and/or newspapers. Each example is labeled with emotions (anger, disgust, fear, joy, sadness, surprise) and polarity (positive/negative). The test data was independently labeled by six anno-

tators. Annotation was performed using a web-based interface that displayed one headline at a time, together with a slide bar for valence assignment. The interval for the valence annotations was set from -100 to 100, where 0 represents a neutral headline, -100 represents a highly negative headline and 100 corresponds to a highly positive headline. We selected only positive or negative sentences, not neutral ones. The inter-annotator agreement for the sentiment polarity is 0.78 (Pearson's correlation).

In order to address the problem mentioned above we proceeded as follows. First of all we specified a training dataset consisting of 36 stimuli—12 positive, 12 negative, and 12 neutral—from behavioral norms (Vinson & Vigliocco 2008; Coltheart, 1981) on whose valence there is substantial agreement among a large number of subjects. Every subject sees each stimuli 5 times. The signal collected from these stimuli is used to train a per-subject valence classifier that is then used to assign a predicted valence to 42 stimuli from the testing dataset (words occurring in a subset of the SemEval test set). The predicted word valences are then fed into a classifier for predicting the overall valence of 12 selected sentences. Our working hypothesis is that the positive, neutral and negative valence of words may be processed by different neural mechanisms and the valence information can be reflected by and extracted from the EEG data. The trained classifier maps the EEG feature space into the negative, neutral and positive valences. Therefore the trained classifier should be able to predict the valence of any test word. Figure 1 sketches out the working procedure described here.

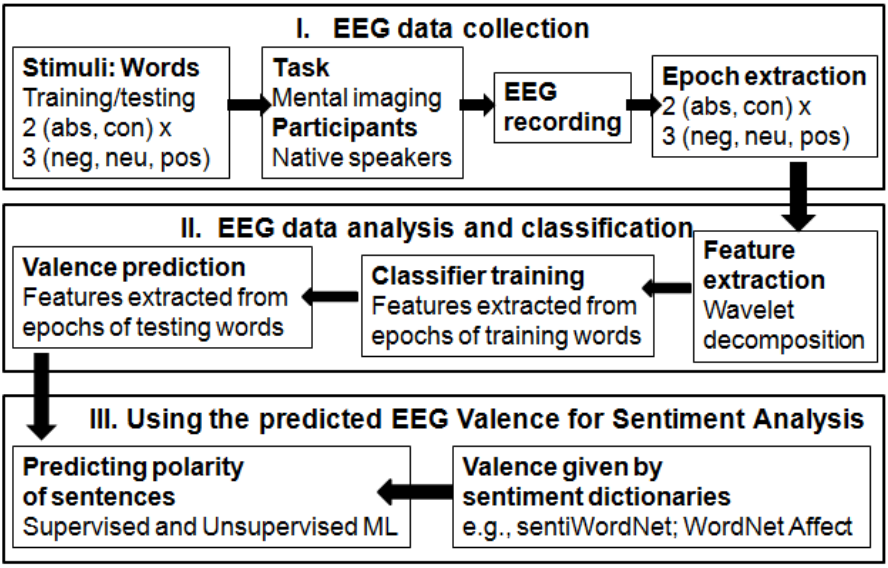

Figure 1: Schematic procedure using brain data for sentiment analysis.

Last but not least, there is the problem of achieving a good performance on determining predicted valence. The performance of EEG at lexical information (Murphy *et al.*, 2011) is typically not comparable to that obtained using fMRI (Mitchell *et al.*, 2008; Pereira *et al.*, 2009). In particular with EEG it is typically more difficult to achieve good inter-subject classification. This can be attributed to the following: 1) the poor spatial resolution of EEG signal; 2) differences in emotional experience between participants. For this reason at present we collect both training and testing data from the same subject.

3 Using machine learning to decode and predict the valence of English words from EEG data

In this Section we discuss how we used EEG to decode the emotional valence of English words.

3.1 EEG experiment and data preprocessing

Materials. Previous work (Kousta et al., 2009; Kousta et al., 2011) suggests that there are likely to be differences with regards to extracting valence between abstract and concrete words. We used therefore a dataset classified according to two dimensions: abstract vs. concrete, or according to their emotional valence (negative, neutral and positive). 36 words were manually selected to vary appropriately in concreteness and valence ratings between the 6 experimental categories and to be otherwise matched in terms of a comprehensive list of linguistic parameters that could serve as confounds. To validate the final set of words, 2-way analysis of variance was undertaken to verify that the experimental groups did not significantly differ in any undesirable way. Results are shown in table 1, where V denotes the main effect was valence category, C denotes the main effect was concreteness category and C×V is the interaction.

Linguistic parameters	V		C		C×V	
	F(1,30)	p	F(2,30)	p	F(2,30)	p
Valence	0.02	0.88	201.26	0	0.89	0.42
Concreteness	266.7	0	0.06	0.93	0.88	0.43
Number of letters	0	1	0	1	0	1
Imageability	84.18	0	0.24	0.79	0.45	0.64
Arousal	0.16	0.7	2.9	0.07	1.35	0.27
Age of acquisition	2.6	0.12	0.25	0.78	0.6	0.56
Familiarity	0.41	0.53	0.58	0.56	1.12	0.34
Log frequency	0	0.99	0.71	0.5	1.22	0.31
Number of orthographic neighbours	0.52	0.47	0.06	0.94	0.15	0.86
Bigram frequency	0.95	0.37	0.1	0.9	0.25	0.78
Number of morphemes	1	0.33	1	0.38	1	0.38

Table 1: Results of 2-way analysis of variance on the training set.

For the test set, we chose 12 sentences from the dataset provided in the SemEval-2007 Sentiment Analysis Task 14 (Strapparava & Mihalcea, 2007) and chose the 42 most frequent

non-stopword nouns. The sentences were chosen in order to have a balance between positive, neutral and negative polarities, as well as between concrete and abstract words. The stimuli in the training set and test set are listed in Table 2. The 12 sentences are listed in Table 3.

Participants. One PhD student and two postdoctoral fellows at the University of Trento took part in the study, all native speakers of English. One participant was male and two female (age range 26–37, mean 33). One identified herself as left-handed, and two as right-handed. All had normal or corrected-to-normal vision. Participants received compensation of €7 per hour. The studies were conducted under the approval of the ethics committee at the University of Trento, and participants gave informed consent.

Training set	Abstract	Negative	harm, hurt, gloom, deceit, terror, sorrow
		Neutral	mood, guess, minute, motive,span, trance
		Positive	cure, ease, peace, reward, warmth, virtue
	Concrete	Negative	jail, scar, blood, corpse, cancer, poison
		Neutral	mule, cart, waist, marble, barrel, cement
		Positive	silk, cash, heart, palace, cherry, silver
Test set	Abstract	save , sick, switch, fetal, loss, swallow, technology, crash, plan, warning, copyright, reject, claim, health, university, offer, support, rabies, suspect, debate, miracle, hail, release, marathon	
	Concrete	Squirrel, boy, park, school, scientist, cocoa, suburb, riot, committee Vaccine, helicopter, river, dolphin, pill, parents, gene	

Table 2: Stimuli in the training and test set.

Number	Sentence	Polarity
1	*Squirrel* jumps *boy* in *park*; *rabies suspected*	-71
2	*University offers support* to New Orleans *school*	+60
3	Beyonce *copyright claim rejected*	-7
4	*Scientists* tout *cocoa*'s *health* benefits	+72
5	*Riot warning* for France *suburbs*	-64
6	*Committee debates* cancer *vaccine plan*	+2
7	Die As US *Helicopter Crashes* in Iraq	-93
8	*Technology* may *save* India's *river dolphins*	+67
9	Poison *Pill* to *Swallow*: Hawks Hurting After *Loss* to Vikes	-35
10	Rescued *boys parents hail* '*miracle*'	+71
11	*Sick* hearts *switch* on a *fetal* gene	-12
12	*Marathon winner released* from *hospital*	+70

Table 3: Test sentences. The words extracted in the test set are highlighted by italic format

Experimental paradigm. Participants saw written words on the screen, repeated 5 times in random order, and are asked to imagine situations exemplifying the words. Once the situation came to mind they responded with a button press. Words were presented until button press, or to a timeout of 5s. Fixations and blanks added 3s per trial. Participants sat in a re-

laxed upright position 60 cm from a computer monitor in reduced lighting conditions. The task duration was split into five blocks and participants were given the choice to pause between each. Each trial began with the presentation of a fixation cross for 0.5 s, followed by the stimulus word, a further fixation cross for 0.5 s and a blank screen for 2 s. Participants were asked to keep still during the task, and to avoid eye-movements and facial muscle activity in particular, except during the 2s blank period.

EEG recording and data preprocessing. The experiment was conducted at the CIMeC/DiSCoF laboratories at University of Trento, using a 64-electrode Brain Vision BrainAmp system, recording at 500 Hz. A wide-coverage montage based on the 10–20 system was used, with a single right earlobe reference, and ground at location AFz. Electrode impedances were generally kept below 10 kOhms. However, sessions including electrodes that exceeded this limit were still included in subsequent analysis, as the techniques used proved robust to such noise. Data preprocessing was conducted using the EEGLAB package (Delorme & Makeig, 2004). The data was band-pass filtered at 1–50 Hz to remove slow drifts in the signal and high-frequency noise, and then down-sampled to 125 Hz. An ICA analysis was next applied using the EEGLAB implementation of the Infomax algorithm (Makeig *et al*., 1996). Artefactual ICA components were then identified and removed by hand in each dataset. Eye-artefact components were removed –usually one component for vertical movements including blinks, and another for horizontal movements.

3.2 EEG data analysis and classification

Wavelet Feature extraction and selection. To classify the EEG data, first of all we extracted data epochs from the preprocessed data in a time window after stimulus onset.

1D multilevel discrete wavelet transform decomposition was employed to extract the decomposition coefficients of the epoched EEG data in the time-frequency domain. Two wavelet functions: coif3 and db7, were used. For a given EEG epoch of a given channel, extracted features were ordered as a list of coefficients arrays in the form [cA_n, cD_n, cD_n-1, ..., cD2, cD1], where n denotes the level of decomposition. The first element (cA_n) of the list is an approximation coefficients array and the following elements (cD_n to cD_1) are details of coefficients arrays. Figure 2 illustrates one EEG epoch of Fpz channel and the extracted wavelet approximation coefficients array and details of coefficients arrays. For a given trial, the extracted EEG features are collected in a wavelet coefficients array whose number of elements equals to the number of channels × the number of coefficients of a single trial in a single channel.

Usually, the number of the extracted features is huge and the feature array contains many redundant or irrelevant features for valence classification. Taking the epoch from 0.1 to 1.4 seconds as an example, the number of the extracted EEG features of each trial is 13568 (= 64 × 212, where 64 is the number of channels and 212 the number of extracted wavelet coefficients). To shorten classifier training time, improve model interpretability and enhance model generalization, we employed univariate ANOVA to select the most promising 3000 features with the highest F-scores.

Classification. **A** SMLR classifier (Krishnapuram *et al*., 2005) was used in 10, 20 and 30 fold cross-validation analyses. The training dataset was constructed by the wavelet features

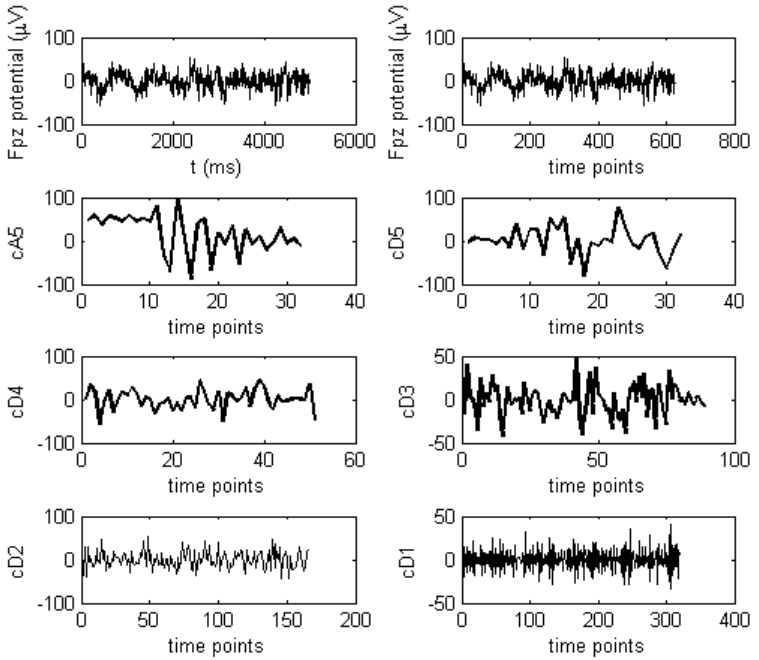

Figure 2: One EEG epoch of Fpz channel and its 5 level wavelet decomposition coefficients.

corresponding to the trials with stimuli in the training set of the words in Table 2. For a given category, abstract (negative, neutral, positive) or concrete (negative, neutral, positive), in the training dataset the total number of samples was 90 (18 words × 5 replicates).

Prediction. The test dataset was constructed by the wavelet features corresponding to the trials with stimuli in the test set of the words in Table 2. The test dataset contained 18× 5=90 concrete words and 24×5=120 abstract words. The test dataset were used as input to the trained classifier to predict the valence of the test words by assigning a valence to each EEG trial with trigger number in the test set.

3.3 Results

In order to get better classification of the emotional valence of English words, we separately classified the valence of concrete and abstract words.

Training the classifier. To train the classifier, for each subject, we tried different time epochs and two wavelet functions coif3 and db7. We found a time period from 0.1 to 1.6 seconds after stimulation onset, in which the classification accuracy is higher. The classification results of the training words are shown in Table 4. Here we show the best classification accuracy for each subject within a time window in the period 0.1 to 1.6 seconds. The

chance to classify into three classes is 33.3%. Our classification accuracy is from 43% to 63%, which is well above chance.

For each concrete and abstract category of each dataset we have also calculated mean classification accuracy over 10 time windows (0.1 or 0.2 to 0.7 + 0.1×n seconds, where n = 0, 1, 2, …, 9). For abstract category, the mean classification accuracy is (43.45±3.62)% for subject 1, (54.44±3.88)% for subject 2 and (40.00±4.67)% for subject 3. For concrete category, the mean classification accuracy is (56.45±3.73)% for subject 1, (52.44±4.81)% for subject 2 and (51.63±6.09)% for subject 3. This result indicates the mean classification accuracies are also well above chance. Especially the three mean classification accuracies of the concrete category are greater than 50%. To study the effect of number of selected features on the classification accuracy, we reduce the number of selected EEG features. We found that for the concrete category, using 300 selected features to train the classifier one can get mean accuracy well above chance. However, for abstract category, in order to get mean accuracy well above chance we have to use 1000 selected features to train the classifier. Therefore we used 1000 selected features for abstract category and from 300 for concrete category to train the classifier. Then we calculated mean classification accuracy over 10 time windows. For abstract category, the mean classification accuracy is (41.20±5.71)% for subject 1, (51.44±4.94)% for subject 2 and (38.49±4.85)% for subject 3. For concrete category, the mean classification accuracy is (42.32±4.67)% for subject 1, (42.17±3.41)% for subject 2 and (48.98±4.03)% for subject 3. This result suggests that the classification accuracy decreases with the number of selected features.

We have randomized the trials of the feature array so that the relationship between the extracted features and the valence label of each trial is randomly matched. We used such random features as input to train the classifier (3000 selected features, 20-fold). Then we calculated the mean classification accuracy over 20 such random EEG for concrete and abstract classes of each dataset in the same epoch as given in Table 4. For abstract category, the mean classification accuracy is (42.94±6.86) % for subject 1, (41.51±5.73)% for subject 2 and (37.98±6.77)% for subject 3. For concrete category, the mean classification accuracy is (42.65±8.61)% for subject 1, (42.34±5.89)% for subject 2 and (41.61±4.88)% for subject 3. The mean accuracy is between (37.98±6.77)% and (42.94±6.86)%. Considering that this result is probably caused by the large number of EEG features, we reduce the number of selected EEG features from 3000 to 1000 for abstract category and from 3000 to 300 for concrete category to train the classifier by the permuted EEG data from 0.1 to 1.6 seconds after stimuli onset. Then we calculated the mean classification accuracy over 20 such permuted EEG for concrete and abstract classes of each dataset. For abstract category, the mean classification accuracy is (37.07±5.26)% for subject 1, (40.11±7.99)% for subject 2 and (36.74±7.08)% for subject 3. For concrete category, the mean classification accuracy is (33.52±9.36)% for subject 1, (36.5±5.77)% for subject 2 and (37.35±6.22)% for subject 3.

Predicting the valence of test words. For each dataset, the classifier trained by the training trials with inside 20-fold training/testing partitions of the data was employed to predict the valence of the words in the test trials. The prediction lists of the abstract and concrete words from the three subjects were employed for sentiment analysis in the following Sec-

tion. Note that for each word there are five trials. Accordingly the classifier predicts five three-way neg-or-neu-or-pos valences for each word.

Subject	Concreteness	Epoch(s)	Wavelet Function	ClassAccuracy (%) (chance = 33.3)
s1	abstract	0.1 to 0.7	db7	47.8 (10 folds); 50.7 (20 folds); 48.9 (30 folds)
	concrete	0.1 to 1.6	db7	46.7 (10 folds); 62.8 (20 folds); 53.3 (30 folds)
s2	abstract	0.1 to 1.4	coif3	54.0 (10 folds); 58.5 (20 folds); 57.8 (30 folds)
	concrete	0.1 to 1.3	coif3	50.0 (10 folds); 57.0 (20 folds); 51.1 (30 folds)
s3	abstract	0.1 to 0.8	coif3	43.3 (10 folds); 51.8 (20 folds); 46.7 (30 folds)
	concrete	0.2 to 1.1	coif3	58.9 (10 folds); 63.0 (20 folds); 57.8 (30 folds)

Table 4: Classification results of the training words.

4 Using EEG valence for sentiment analysis

In this Section we discuss how the valences extracted from EEG were good predictors of the sentiment polarity of the 12 selected sentences, using machine learning techniques.

4.1 Comparison with existing resources and supervised sentiment analysis

After collecting brain data for 3 native English subjects, we had 5 trials for each word as integer numerical features, and we exploited them for machine learning. We wanted to predict sentence polarities and compare the results to the predictions derived using word polarities from two different lexical resources: SentiWordNet[1] (Baccianella *et al*., 2010; Esuli & Sebastiani, 2006) and SenticNet[2] (Cambria *et al*., 2012). The classification task is binary, as the target class to predict is sentence polarity (positive/negative), given as features the positive, negative and neutral word polarities from the EEG signal in the first case and from the lexical resources in the second one.

Subject performance comparison. As for the first experiment, we tested different algorithms and compared the classification performance of the three subjects in order to identify the best one. We used as features the sum of the brain values and as target class the sentence polarity (positive/negative), using 3-fold cross validation as evaluation setting in Weka (Witten & Frank, 2005). Results, reported in Table 5, show that there is not a single algorithm that works best. Among the subjects, Subject 3 achieved the best performance either on concrete and abstract words, using a Sequential Minimal Optimization (Platt, 1998) algorithm. We used the best performing subject (subject 3) to select the best method to use the 5 trial values for the classification task.

1 http://sentiwordnet.isti.cnr.it/

2 http://sentic.net/

Feature selection: all trials vs. sum of values. We ran an experiment to test how the different brain outcomes in the 5 trials can be exploited to achieve the best results. In one test we used all the 5 trials as features, while in the second test we exploited the sum of the values —which can be +1, -1 and 0— as one feature. As before, we used a 3-fold cross validation in Weka. The result, computed using SMO and averaged over the three subjects and over abstact and concrete words, are f1=0.442 using all the values, and f1=0.407 using the sum of trials.

Comparing brain data and lexical resources. Then we extracted from SentiWordNet and SenticNet all the values associated to the selected words, leaving a tie if no values were available. We had 14 ties with SenticNet and no ties with SentiWordNet. SenticNet provides one polarity value (positive or negative), while SentiWordNet provides one value for the positive pole and one for the negative one. Polarities from SentiWordNet have been extracted from the first sense; if both positive and negative values were available, we used the difference between the two.

Data	Concreteness	Algorithm	Precision	Recall	F1measure
baseline	abstract	zeroRule	0.25	0.5	0.333
s1		SMO	0.349	0.375	0.347
s2		bayes	0.594	0.583	0.571
s3		SMO	*0.752*	*0.708*	*0.695*
senticNet		SMO	0.757	0.75	0.748
SentiWN		logistic	**0.853**	**0.792**	**0.782**
baseline	concrete	zeroRule	0.309	0.556	0.397
s1		logistic	0.494	0.5	0.495
s2		bayes	0.444	0.444	0.444
s3		SMO	***0.797***	***0.778***	***0.778***
SenticNet		logistic	0.728	0.722	0.723
SentiWN		SMO	0.477	0.5	0.475

Table 5: Comparison of supervised analysis results obtained by brain data and dictionaries.

Like before, we ran the experiment using 3-fold cross validation in Weka to predict the polarity of sentences. Results, reported in Table 5, show that lexical resources yield better classification performances for abstract words, but also that subject 3 achieved the best performance on concrete words. The correlation coefficients are r = 0.648 for subject 3 with concrete words and r = 0.345 with SentiWordNet on abstract words.

4.2 Integrating the valence in a state-of-the-art unsupervised sentiment analysis system

For the unsupervised scenario we used the sentiment analyser (Steinberger *et al.*, 2011) developed as part of the Europe Media Monitor (Atkinson & Van der Goot, 2009). The objective of the analyser is to detect positive or negative opinions expressed towards entities in the news across different languages and to follow trends over time.

It attaches a sentiment score to all entity mentions, mainly persons and organizations. It uses a fixed window of 6 terms, which was found to be optimal in the analysis in Balahur *et al.*, 2010, around the entity mention to look for sentiment terms. The approach also accounts for contextual valence shifting (negations, diminishers and intensifiers). In their case, the approach is rather defensive, as it looks for shifters only two terms around each sentiment term. This way it captures the most common shifters (very good, not good, less good) but modals or adverbs with larger scope may not be captured. For our purpose the tool was modified to analyze the whole sentence regardless an entity mention and regardless any fixed window for sentiment terms.

The approach uses language-specific sentiment dictionaries. Inspired by the positive effect of introducing two levels of sentiment intensity in Balahur *et al.*, 2010, it uses more classes. The score of positive terms is 2, negative -2, very positive 4, and very negative -4. If a polar expression is negated, its polarity score is simply inverted. In the case of term with higher intensity we lower the intensity. In a similar fashion, diminishers are taken into consideration. The difference is, however, that the score is only reduced rather than shifted to the other polarity type. Special care has to be taken when shifters are combined: for example not very good – good carries the score (+2), it is intensified by very (+3) and inverted by not, however, if we take the same approach as in the case of optimal above, the result is (-2). The scores of the sentiment terms found in a sentence are summed up and the normalized score gives the final sentiment of the sentence. The score ranges from -100 to +100, where, for instance, 100 corresponds to a case with all the terms very positive. The score thus corresponds to the range of SemEval-2007.

Sentiment Dictionaries. We tested the following resources:

- WordNet Affect (WNA) (Strapparava & Valitutti, 2004): categories of anger and disgust were grouped under high negative, fear and sadness were considered negative, joy was taken as containing positive words and surprise as highly positive.
- SentiWordNet (SWN) (Esuli & Sebastiani, 2006): we used the difference between the positive and negative scores. We mapped the positive scores lower than 0.75 to the positive category, the scores higher than 0.75 to the highly positive set, the negative scores lower than 0.75 to the negative category and the ones higher than 0.75 to the highly negative set.
- MicroWordNet (MWN) (Cerini *et al.*, 2007): the mapping was similar to SentiWordNet.
- General Inquirer (GI) (Stone *et al.*, 1966): besides other annotations, each English word is labeled as “positive outlook” or “negative outlook” in GI. Terms taken from these categories formed one of the first sentiment dictionaries.

- JRC dictionaries (JRC) (Steinberger *et al*., 2012): semi-automatically collected subjective terms in 15 languages. Pivot language dictionaries (English and Spanish) were first manually created and then projected to other languages. The 3rd language dictionaries were formed by the overlap of the translations (triangulation). The lists were then manually filtered and expanded, either by other relevant terms or by their morphological variants, to gain a wider coverage.

We run the analyser on the 12 sentences selected from the SemEval-2007 corpus. We used the above mentioned dictionaries, including the brain data. The results are shown in Table 6.

Data	Precision	Recall	F1 measure
s1-abs	0.556	0.238	0.333
s1-conc	0.833	0.238	0.37
s2-abs	0.444	0.19	0.267
s2-con	0.714	0.238	0.357
s3-abs	0.333	0.143	0.2
s3-con	0.778	0.333	0.467
JRC	**1**	**0.619**	**0.765**
GI	0.923	0.571	0.706
SWN	0.706	0.571	0.632
WNA	0.524	0.524	0.524
MWN	0.625	0.238	0.345

Table 6: Comparison of unsupervised analysis results obtained by brain data and various dictionaries.

In the case of using the JRC dictionary, all system judgments were correct or the system did not find any sentiment term resulting in a recall error. This corresponds to the fact that the system was developed to be precision-oriented. The correlation coefficient was r=0.688. Precision values achieved by subjects on concrete words outperform precision of WordNet-Affect, sentiWordNet and Micro-WordNet. With the s3-con dictionary the correlation coefficient was r=0.254.

However, the performance of recall of human subjects is worse than the lexical resources, and this influences the final f1-measure. In general, the supervised approaches perform better, as they can work with more information than the simple presence/absence of a word and there is the learning phase.

5 Conclusions

In this paper we report exploratory experiments testing whether text valence can be reliably extracted from brain signals using EEG—at present, the only technology that can be expected to be usable to elicit brain information on a large scale, in particular when the new generation of low-cost headsets will appear. Our results demonstrated that the emotional valence information of words can indeed be extracted by wavelet decomposition coefficients and classified by machine learning with accuracy well above chance.

We also carried out very preliminary experiments using lexical valence extracted from EEG for sentiment analysis of a small set of sentences from a standard dataset, using both supervised and unsupervised machine learning techniques. For those sentences at least, the precision achieved using lexical valence extracted from EEG is close to the one obtained using standard sentiment dictionaries such as WordNet Affect, senticNet or SentiWordNet. EEG-based sentiment analysis results are even better when using supervised learning. We conclude that the paradigm we propose might indeed develop into an alternative technique for collecting valence.

Our next step will be to test these methods on a larger scale, in three respects. First of all, we started to use larger datasets of sentences from the sentiment analysis shared task at SemEval-2013; and to test our methods on Italian as well as English. Second, we started to also use adjectives, adverbs and verbs as stimuli. Last but not least, we started to investigate the effect of context on the valence of words such as *rude* that have a negative valence in sentences such as *You're being rude* but a positive one in sentences such as *I found him in rude health*. We intend to study how the valences of emotional words are modified by different contexts and how their emotional categories change with contexts. We are also interested in investigating how the emotional words and emotional mood exert influence on sentence processing and on the polarity of sentences, as it has been recently found that emotional valence in a word and emotional mood of the participants inducted by film clips impact the syntactic and semantic processing (Chwilla *et al*., 2011; Martín-Loeches, *et al*., 2012). From a methodological perspective, we aim to improve the classification accuracy by selecting most informative channels and extracting other EEG features such as event-related potential and the reconstructed wavelet approximation and details of the EEG data.

Acknowledgements

This research was carried out as part of the Deep Relations project, a collaboration between CIMEC, Expert Systems and Fondazione Bruno Kessler (FBK) funded by the Provincia di Trento. It was also partly supported by projects: NTIS (Net Technologies for Information Society), European Center of Excellence, CZ.1.05/1.1.00/02.0090 and MediaGist. EU's FP7 People Programme (Marie Curie Actions), nº 630786.

References

Adorni, R., Proverbio, A.M. (2009). New insights into name category-related effects: is the Age of Acquisition a possible factor? Behav Brain Funct 5: 33.

Artstein R., Poesio M.. (2008). Intercoder agreement for Computational Linguistics. In Computational Linguistics, 34(4): 555—596.

Atkinson, M. and Van der Goot, E. (2009). Near Real Time Information Mining in Multilingual News. In 18th International World Wide Web Conference, WWW.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC(10): 2200-2204.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. In Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC.

Cambria, E., Havasi, C., and Hussain, A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In FLAIRS Conference 202-207.

Cato, M.A., Crosson, B., Gökcay, D., Soltysik, D., Wierenga, C., Gopinath, K., Himes, N., Belanger, H., Bauer, R.M., Fischler, I.S., Gonzalez-Rothi, L., & Briggs, R.W. (2004). Processing Words with Emotional Connotation: An fMRI Study of Time Course and Laterality in Rostral Frontal and Retrosplenial Cortices. Journal of Cognitive Neuroscience 16(2): 167–177.

Cerini, S., Compagnoni, V., Demontis, A. (2007). Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Italy.

Chwilla, D. J., Virgillito, D., & Vissers, C. T. W. M. (2011). The relationship of language and emotion: N400 support for an embodied view of language comprehension. Journal of Cognitive Neuroscience 23(9): 2400–2414.

Coltheart, M. (1981). The MRC psycholinguistic database. In Quarterly Journal of Experimental Psychology. 33(A): 497–505.

Craggs, R. & McGee Wood, M. (2004). A two dimensional annotation scheme for dialogue. In Proc. Of AAAI Spring Symposium.

Crosson, B., Cato, M. A., Sadek, J., Radonovich, K., Go¨kc¸ay, D., Bauer, R., Fischler, I., Maron, L., Auerbach, E., Browd, S., Freeman, A., & Briggs, R. (2002). Semantic monitoring of words with emotional connotation during fMRI: Contribution of left-hemisphere limbic association cortex. Journal of the International Neuropsychological Society 8: 607–622.

Crosson, B., Radonovich, K., Sadek, J. R., Go¨kc¸ay, D., Bauer, R. M., Fischler, I. S., Cato, M. A., Maron, L., Auerbach, E. J., Browd, S. R., & Briggs, R. W. (1999). Left-hemisphere processing of emotional connotation during word generation. NeuroReport 10: 2449–2455.

Delong, K. A., Urbach, T. P., & Kutas, M. M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience 8(8): 1117–1121.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134(1): 9-21.

Esuli, A. & Sebastiani, F. (2006). Determining Term Subjectivity and Term Orientation for Opinion Mining. In Proceedings of EACL2006 193-200.

Fuggetta, G., Rizzo, S., Pobric, G., Lavidor, M., & Walsh, V. (2009). Functional representation of living and nonliving domains across the cerebral hemispheres: a combined event-related potential/transcranial magnetic stimulation study. J Cogn Neurosci 21: 403–414.

Gu, Y., Cazzolli, G., Murphy, B., Miceli, G, & Poesio, M. (2013). EEG study of the neural representation and classification of semantic categories of animals vs tools in young and elderly participants. BMC Neuroscience 14 (Suppl 1): 318 http://www.biomedcentral.com/bmcneurosci/supplements/14/S1

Hagoort, P., Hald, L., Bastiaansen, M. C. M., & Petersson, K.-M. (2004). Integration of word meaning and world knowledge in language comprehension. Science 304(5669): 438–441.

Hirata, M., Koreeda, S., Sakihara, K., Kato, A., Yoshimine, T., & Yorifuji, S. (2007). Effects of the emotional connotations in words on the frontal areas—a spatially filtered MEG study. Neuroimage 35(1): 420-429.

Hoenig, K., Sim, E.J., Bochev, V., Herrnberger, B., Kiefer, M. (2008). Conceptual flexibility in the human brain: dynamic recruitment of semantic maps from visual, motor, and motion-related areas. J Cogn Neurosci 20: 1799–1814.

Kiefer, M. (2001). Perceptual and semantic sources of category-specific effects: event-related potentials during picture and word categorization. Mem Cognit 29: 100–116.

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. Cognition 112(3): 473-481.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. Journal of Experimental Psychology: General 140(1): 14.

Krishnapuram B., Carin, L., Figueiredo, M.A.T. & Hartemink, A.J. (2005). Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6): 957-9368.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Makeig, S., Bell, A. J., Jung, T. P., and Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. Advances in neural information processing systems 145-151.

Martín-Loeches, M., Fernández, A., Schacht, A., Sommer, W., Casado, P., Jiménez-Ortega, L., & Fondevila, S. (2012). The influence of emotional words on sentence processing: electrophysiological and behavioral evidence. Neuropsychologia 50(14): 3262–3272.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science 320(5880): 1191-1195.

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., and Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. Brain and language, 117(1): 12-22.

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

Paz-Caballero, D, Cuetos, F, & Dobarro, A. (2006). Electrophysiological evidence for a natural/artifactual dissociation. Brain Res 1067: 189–200.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45(1): S199-S209.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In B. Schoelkopf and C. Burges and A. Smola, (editors), Advances in Kernel Methods - Support Vector Learning.

Proverbio, A.M., Del Zotto, M. & Zani, A. (2007). The emergence of semantic categorization in early visual processing: ERP indices of animal vs. artifact recognition. BMC Neurosci 8: 24.

Pulvermüller, F., Lutzenberger, W. & Preissl, H. (1999). Nouns and verbs in the intact brain: evidence from event-related potentials and high-frequency cortical responses. Cereb Cortex 9: 497–506.

Renoult, L., & Debruille, J. B. (2010). N400-like potentials and reaction times index semantic relations between highly repeated individual words. J Cogn Neurosci 23(4): 905–922.

Renoult, L., Davidson, P. S. R., Palombo, D. J., Moscovitch, M., & Levine, B. (2012). Personal semantics: at the crossroads of semantic and episodic memory. Trends Cogn Sci 16(11): 550–558.

Simanova, I., van Gerven, M., Oostenveld, R. & Hagoort, P. (2010). Identifying Object Categories from Event-Related EEG: Toward Decoding of Conceptual Representations. PloS one 5(12): e14465. doi:10.1371/journal.pone.0014465.

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R. and Van der Goot, E. (2011). Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In Proceedings of the 8th International Conference Recent Advances in Natural Language Processing 770-775. Hissar, Bulgaria.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S. & Zavarella, V. (2012). Creating sentiment dictionaries via triangulation. In Decision Support Systems (53): 689–694, Elsevier.

Stone, P., Dumphy, D., Smith, M., Ogilvie, D. (1996). The general inquirer: a computer approach to content analysis. M.I.T. Studies in Comparative Politics. M.I.T. Press, Cambridge, MA.

Strapparava, C. & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In Proceedings of the 4th International Workshop on the Semantic Evaluations, Prague, Czech Republic.

Strapparava, C. & Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. Behavior Research Methods 40(1): 183-190.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.