Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models

Elisavet Palogiannidi[1]*, Elias Iosif* [2]*, Polychronis Koutsakis*[1]*, Alexandros Potamianos*[2,3]

[1]School of ECE, Technical University of Crete, Chania 73100, Greece
[2]"Athena" Research and Innovation Center, Maroussi 15125, Athens, Greece
[3]School of ECE, National Technical University of Athens, Zografou 15780, Greece
epalogiannidi@isc.tuc.gr, {iosife,potam}@central.ntua.gr, polk@telecom.tuc.gr

Abstract

We propose and evaluate the use of an affective-semantic model to expand the affective lexica of German, Greek, English, Spanish and Portuguese. Motivated by the assumption that semantic similarity implies affective similarity, we use word level semantic similarity scores as semantic features to estimate their corresponding affective scores. Various context-based semantic similarity metrics are investigated using contextual features that include both words and character n-grams. The model produces continuous affective ratings in three dimensions (valence, arousal and dominance) for all five languages, achieving consistent performance. We achieve classification accuracy (valence polarity task) between 85% and 91% for all five languages. For morphologically rich languages the proposed use of character n-grams is shown to improve performance.

Index Terms: affective lexicon, affective ratings, valence, arousal, dominance, semantic similarity, sentiment analysis, emotion recognition

1. Introduction

Emotion is mainly conveyed by speech, but it can still be perceived in text and it can be elicited by its content and form [1]. Readers experience emotion by placing themselves in the position of characters of a narrative and imagining their own emotional reaction [2]. Affective text analysis has recently attracted much interest from the research community [3–6]. An open research question is whether affective analysis of text can be handled equally well across different languages. Affective analysis for multiple languages was investigated in [7] for subjectivity detection and in [8] for polarity (valence) prediction. Computational tools of affective text analysis typically rely on the exploitation of affective lexica. Affective lexica consist of word entries (usually about 1K) of a target language, that are annotated with respect to emotional dimensions, usually valence, arousal and dominance. Examples regarding the manual creation of lexica are available in the literature for languages such as English [9], Spanish [10], European Portuguese [11] and Greek [12]. Since the size of such manually created resources is limited, the development of models that automatically expand them is very important. For example, an affective resource for Russian and Romanian was automatically created in [4]. The availability of multilingual affective resources allows the investigation of the universality of text-based affective models for different languages, enabling the development of cross-language tools.

Textual affective models have been applied to numerous problems such as sentiment analysis and opinion mining [13, 14], affective analysis of social media [15–17], product reviews [18], news headlines [3], emotion prediction on spoken dialogues [19, 20]. Affective analysis of text is applied on various lexical units, such as words [21–23], phrases [6, 24], sentences [23], even whole documents [25]. Affective analysis of text can contribute to the development of multimodal affective systems [26–28].

Features choice is at the core of affective computational models and often depends on the target application. The simplest textual features used in most affective models employ the lexical information itself. In [29], the count of words is used for personality recognition. Non-lexical features such as punctuation marks, emoticons or references to other users [17] are popular for the affective analysis of user-created content in social media. Semantic features, have also been employed in affective models [5, 23], based on the assumption that "*semantic similarity can be translated into affective similarity*". Semantic similarities can be computed from a corpus using co-occurrence- or context-based metrics. Context-based semantic similarities [30] are motivated by the hypothesis that "*similarity of context implies similarity of meaning*" [31].

An affective-semantic model that automatically expands small affective lexica is of great relevance for the creation of lexica with good vocabulary coverage. In this paper, we apply a model that was first proposed in [23] in similar fashion to [6]. The model is expanded to model more affective dimensions (valence, arousal, dominance) and languages (English, German, Greek, Portuguese and Spanish). Additional context-based semantic similarity metrics are explored using both word and character n-grams as contextual features, and early fusion schemes are being investigated. In addition, alternative criteria (ridge regression) are investigated for training the model parameters. The proposed model is evaluated for each of the aforementioned languages using as ground truth the human scores provided for each dimension.

2. Affective Model

As illustrated in Figure 1, the model takes as input a small affective lexicon and consists of two modules: the semantic similarity computation module (semantic features) and the mapping from the semantic to an affective space (semantic-affective map). A small manually annotated affective lexicon is required for bootstrapping the process. Specifically, the affective model is fed with a number of seed words (a subset of the affective lexicon) and their corresponding affective ratings, in order to train the semantic-affective map. The affective rating of unknown words is then estimated based on the assumption that words that

are semantically related are also affectively related, leading to an expanded affective lexicon.

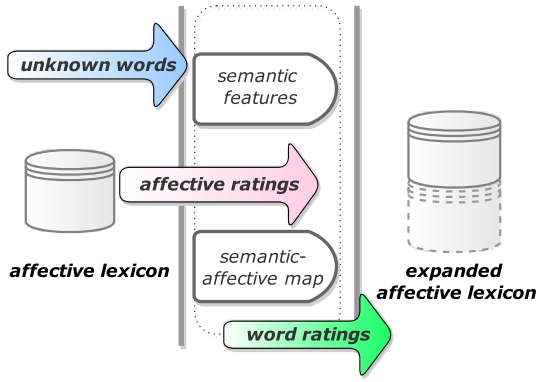

Figure 1: Abstract representation of the affective model.

According to [23], the semantic-affective model estimates affective ratings as a weighted linear combination of semantic similarities between the unknown and the seed words, as follows:

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i v(w_i) S(w_j, w_i), \quad (1)$$

where w_j is the unknown word, $w_{1..N}$ are the seed words, $v(w_i)$, a_i are the affective rating and the weight corresponding to the word w_i and $S(\cdot)$ is the semantic similarity metric between two words. An important advantage of the proposed affective model is the fact that it requires only a small affective lexicon in order to estimate affective ratings for any number of unknown words. Next we detail how the seed word weights a_i and the semantic similarity metrics are estimated.

2.1. Estimation of weights

The a_i weights are used in (1) because not all seed words are equally salient for the estimation of affective ratings. Supervised learning can be employed for estimating these weights as follows:

$$\boldsymbol{X\beta} = \boldsymbol{y}, \quad (2)$$

where $\boldsymbol{X}$ is a $\mu \times N$ matrix containing μ training samples and N features for each sample. β is a $N \times 1$ vector including the a_i weights, while y is a $N \times 1$ vector containing the known affective ratings. According to *Least Squares Estimation* (LSE), the weights can be estimated as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 \quad (3)$$

LSE may yield weights with large variance, so we investigate the use of Ridge Regression for alleviating this problem. *Ridge Regression* (RR) uses the estimator shown in (4), incorporating a regularization factor, λ, which forces the weights to shrink toward zero.

$$\hat{\boldsymbol{\beta'}} = \underset{\beta'}{\operatorname{argmin}} \left[||\boldsymbol{y} - \boldsymbol{X\beta'}||^2 + \lambda ||\boldsymbol{\beta'}|| \right], \quad (4)$$

where β' is the weights vector. The λ values should be greater than zero, while for $\lambda = 0$ the LSE and RR estimators are identical [32].

2.2. Semantic Features

The $S(\cdot)$ metric used in (1) can be implemented within the framework of distributional semantic models (i.e., coprus-based models) that are based on the assumption that words that occur in similar context tend to be semantically related [30, 33]. Having a vocabulary and a corpus in a target language, we can create a contextual feature vector for each vocabulary entry w_i, as follows. Lexical features are extracted after centering a contextual window of size $2H+1$ words on each instance of the target word w_i in the corpus. Then, a feature vector x_i is formulated by extracting the words in distance H from the window center. The semantic similarity between two words w_i, w_j can be computed as the cosine of their respective feature vectors:

$$Q^H(w_i, w_j) = \frac{x_i . x_j}{||x_i||\ ||x_j||}. \quad (5)$$

The aforementioned feature extraction typically deals with words. However, it can be modified by extracting the character n-grams from the words that are captured by the applied contextual window. Usually, the selection of n depends on the desired amount of lexical information that n-grams carry. The elements of the feature vectors are set according to two schemes: 1) a weighting scheme based on the frequency of the features, 2) a binary scheme (B), where each element is set to one if the corresponding feature frequency is at least one and zero otherwise [34]. The weighting scheme we use is based on the point-wise mutual information (PMI). The PMI between a word w_i and the k–th feature of its vector x_i, f_i^k, is computed as shown in (6) [35].

$$\mathrm{PMI}(w_i, f_i^k) = -\log \frac{\hat{p}(w_i, f_i^k)}{\hat{p}(w_i)\hat{p}(f_i^k)}, \quad (6)$$

where $\hat{p}(w_i)$ and $\hat{p}(f_i^k)$ are the occurrence probabilities of w_i and f_i^k , respectively, while the probability of their co-occurrence (within the H window size) is denoted by $\hat{p}(w_i, f_i^k)$. The corpus-based frequencies of lexical items (words or character n-grams) were used in order to compute the probabilities, according to maximum likelihood estimation. The scores derived using PMI lie in the $[-\infty, +\infty]$ interval. In particular, we use the positive point-wise mutual information (PPMI), in order to bound the computed scores within the $[0, +\inf]$ interval. PPMI is a special case of PMI according to which the negative PMI scores are set to zero, based on the assumption that the contextual features that exhibit negative PMI do not contribute to the estimation of similarity much [36].

3. Experimental Procedure

The affective-semantic model is used for the *affective lexicon expansion* task for five languages, namely English, German, Greek, Portuguese and Spanish. This is done for all three affective dimensions valence, arousal and dominance (V,A,D). The English affective lexicon, a.k.a. ANEW [9] contains 1034 words rated in the three continuous affective dimensions (V,A,D). ANEW words were translated into Spanish [10], European Portuguese [11] and Greek [12] and rated by native speakers on the same affective dimensions. The German affective lexicon consists of 2902 words rated on valence and arousal. To simplify performance comparisons we selected a subset with 1034 words and similar ratings distributions with the rest of the lexica.

A corpus per target language was harvested for the computation of context-based semantic similarities. The corpora we

used were created using web data as follows. The process starts with the definition of a vocabulary for each language: 135K, 332K, 407K, 125K, 187K entries for English, German, Greek, Portuguese and Spanish, respectively. For each word of the vocabulary a web search query was formulated and the snippets of 1K top-ranked documents were downloaded and aggregated.

In order to show the impact of the semantic features on each language's affective lexicon expansion, we employ the context-based semantic similarity metric. The contextual window size is set to $H = 1$ and variations of the contextual features and weighting schemes are used in order to create the similarity metrics described below. The metrics employ two types of contextual features, namely, words (W) and character n-grams (ngram). Two types of weighting schemes are also employed as described in Section 2. Early fusion schemes were also investigated, combining feature vectors that consist of different size n-grams, or of n-grams and words. In Table 1 we list the contextual features and the weighting schemes that are employed by each context-based similarity metric.

Similarity metric	*Contextual features*		*Weigh. scheme*	
	Words	*Character n-grams*	*PPMI*	*Binary*
W-B	✓	×	×	✓
W-PPMI	✓	×	✓	×
4gram-B	×	$n = 4$	×	✓
4gram-PPMI	×	$n = 4$	✓	×
2/3/4/gram-B	×	$n = 2, 3, 4$	×	✓
W+4gram-PPMI	✓	$n = 4$	✓	×

Table 1: Contextual features and weighting schemes for the semantic similarity metrics.

Moreover, the RR estimator was implemented for estimating the weights used in (1). RR requires a tuning step in order to find the appropriate value λ, that maximizes the affective model's classification accuracy. The tuning step is repeated for each language and each affective dimension, using the W-PPMI semantic similarity metric using held-out data.

4. Results

In this section, we present and discuss the performance of the affective lexicon expansion for the various languages and affective dimensions. Results are presented using different semantic similarity computation methods (contextual features and weighting schemes) and weight estimation approaches. For affective model evaluation, we use the seed selection algorithm of [23] and we apply 10-fold cross validation using the affective lexicon of each language both for parameter estimation and for evaluation. An important parameter in our experiments is the number of seeds N in (1). In each fold 10% of the affective lexicon words is used as test (unknown words) and 90% is used as train. The model performance is captured through binary classification accuracy (positive vs. negative values) and Pearson correlation between the automatically estimated and the manually annotated valence ratings.

The performance for each affective dimension and language as a function of the seed words, using the W-PPMI semantic similarity metric is shown in Figure 2. The affective model appears to be robust and well-performing when at least 100 seeds are used. The highest performance is reached for 500-600 seeds, for all the affective dimensions. Performance for valence (Figure 2(a), 2(d)) is consistent across all languages while the highest scores are obtained for English and Portuguese. Performance is consistent also for arousal (Figure 2(b), 2(e)) and dominance (Figure 2(c), 2(f))[1]. Among the three affective dimensions, the highest performance is achieved for valence, while the poorest performance is achieved for arousal[2]. The reasons for the differences in the results across languages are not easily interpreted, however they could be attributed to language morphology differences.

The correlation and classification accuracy of the affective (valence) lexicon expansion across the different languages is shown in Table 2. The results are presented for the semantic similarity metrics described in Section 3, using 600 seeds and LSE[3]. Although the performance is consistently high for all languages and semantic similarity metrics, minor differences exist between them. The PPMI weighting scheme is superior to binary in all cases, especially for German. The highest performance is achieved for English and European Portuguese when the feature vector consists of the contextual words. The use of context character n-grams yields a slight improvement to performance especially for morphologically rich languages when the binary scheme is used. The early fusion scheme that uses both the contextual words and 4-grams achieves the highest performance for almost all languages.

As seen in Figure 2, the performance of the model is not always robust for a large number of seeds. We investigated whether the low performance of arousal for Spanish and Greek can be attributed to the weights estimation method. For this purpose, we use RR with $\lambda = 0.05$ (derived after tuning the parameter for both languages on arousal, W-PPMI and 600 seeds). The classification accuracy and the correlation of the arousal ratings for Spanish and Greek (the two languages with the lowest performance in arousal) using LSE and RR for 10 up to 900 seeds are shown in Figure 3(left) and Figure 3(right), respectively.

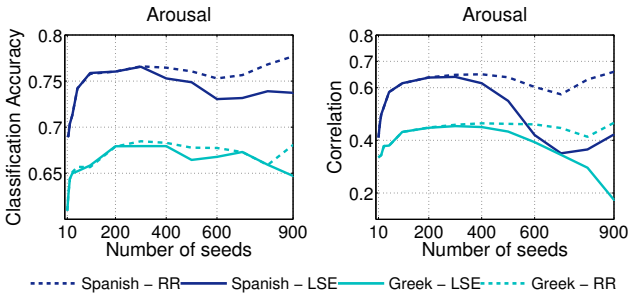

Figure 3: Classification accuracy (left) and correlation (right) for arousal using LSE and RR for the estimation of weights.

It is observed that, as the number of seeds increases the arousal model that uses weights estimated with RR becomes superior to the model that uses weights estimated with LSE. The model that employs RR is robust to a large number of seeds compared to LSE. Additionally, we observed that the best performance achieved using 600 seeds, W-PPMI and LSE can be improved

[1] Dominance ratings are not available for German.

[2] Valence and dominance are highly correlated in all four languages.

[3] In addition to the reported similarity metrics, W-normalized PPMI was also investigated (with no significant performance improvement), as well as 2grams-B and 3grams-B (achieving lower performance than 4grams-B).

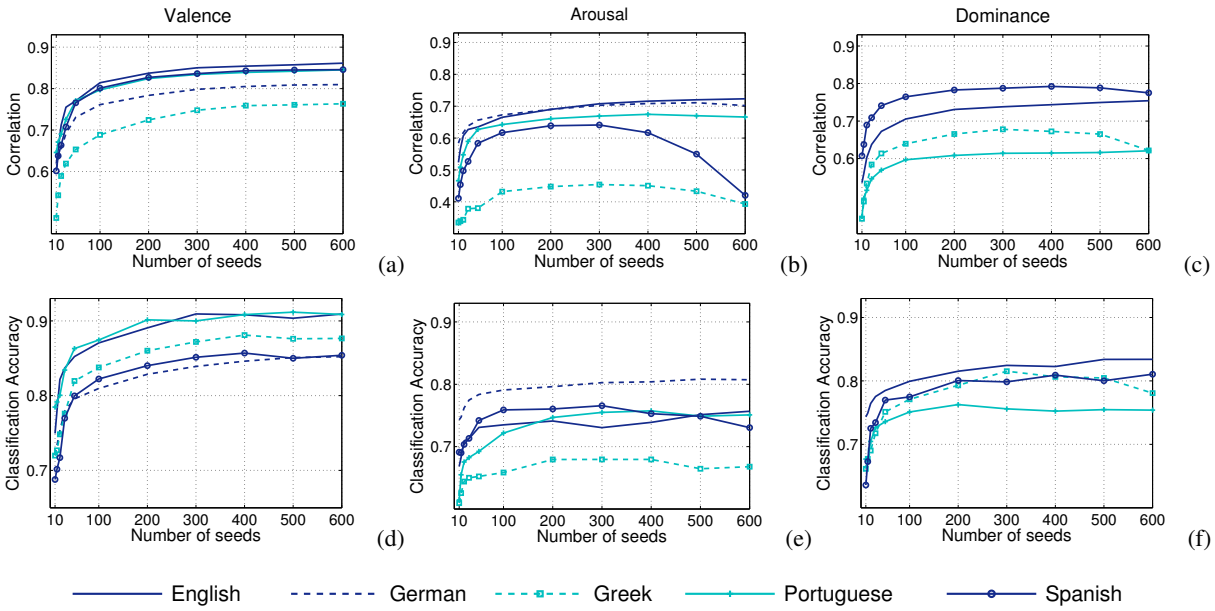

Figure 2: Performance of the affective-semantic model in terms of correlation (a, b, c) and classification accuracy (d, e, f) for valence (a, d), arousal (b, e) and dominance (c, f) across all five languages, using W-PPMI for similarity computation.

Language / Similarity Metric	English		Greek		Spanish		Portuguese		German	
	PC	CA(%)	PC	CA(%)	PC	CA(%)	PC	CA(%)	PC	CA(%)
W-B	0.80	86.9	0.74	84.3	0.84	85.9	0.82	89.3	0.68	77.1
W-PPMI	**0.86**	**90.9**	0.76	87.6	0.84	85.3	**0.84**	**90.8**	**0.80**	**85.2**
4gram-B	0.82	87.8	0.77	87.8	0.84	86.4	0.80	87.6	0.78	82.3
4gram-PPMI	0.84	89.8	**0.78**	**87.5**	**0.85**	**87.7**	0.82	87.4	**0.80**	82.6
2/3/4gram-B	0.82	88.1	0.75	87.6	0.83	86.6	0.80	86.4	0.78	82.2
W+4gram-PPMI	**0.85**	**90.5**	**0.79**	**87.2**	**0.85**	**87.9**	**0.83**	89.3	**0.80**	83.0

Table 2: Classification Accuracy (CA) and Pearson Correlation (PC) between the manually rated and the automatically estimated valence scores for 600 seeds, across the five languages using various semantic similarity metrics.

(up to 0.5-1%) for almost all languages when 900 seeds and RR are used.

5. Conclusions

In this work, we expanded the affective lexica of five languages, namely, English, German, Greek, Portuguese and Spanish for the three affective dimensions valence, arousal and dominance utilizing semantic models. Our approach was found to be applicable across all languages and affective dimensions. Minor differences in performance could be attributed to the linguistic properties of each language e.g., morphology. We investigated various parameters for the context-based computation of semantic similarity observing that the character 4-grams (extracted from the contextual words) are salient features for this task. We showed that the performance of the affective model depends on the weights estimation method, especially for a large number of seeds. Specifically, the employment of a more robust approach for the estimation of weights such as Ridge regression leads to small performance improvements.

In future work we plan to model the role of morphology in the computation of semantic similarity. Also, we aim to verify the universality of our findings by experimenting with more languages. Last but not least, we aim to investigate the compositional aspects of emotion, i.e., how the affective content of words can be composed in order to estimate affective scores for larger lexical units such as phrases and sentences.

Acknowledgements This work has been been partially funded by the SpeDial project ("Machine-Aided Methods for Spoken Dialogue System Enhancement and Customization for Call-Center Applications") supported by the EU Seventh Framework Programme (FP7), grant number 611396.

6. References

[1] R. Picard, *Affective computing*. MIT press, 2000.

[2] D. Keltner and P. Ekman, "Introduction to expression of emotion," in *R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), Handbook of Aective Sciences*. Oxford University Press, 2003, pp. 411–414.

[3] C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective text," in *Proc. of SemEval*, 2007, pp. 70–74.

[4] V. Bobicev, V. Maxim, T. Prodan, N. Burciu, and V. Angheluş, "Emotions in words: Developing a multilingual WordNet-

Affect," in *Computational Linguistics and Intelligent Text Processing*, 2010, pp. 375–384.

[5] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Kernel models for affective lexicon creation." in *Proc. of Interspeech*, 2011, pp. 2977–2980.

[6] P. Turney and M. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus, technical report ERC-1094 (NRC 44929)." National Research Council of Canada, Tech. Rep., 2002.

[7] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *Proc. of Empirical Methods in Natural Language Processing*, 2008, pp. 127–135.

[8] Z. Kozareva, "Multilingual affect polarity and valence prediction in metaphor-rich texts." in *Proc. of Association for Computational Linguistics*, 2013, pp. 682–691.

[9] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings, technical report c-1," The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[10] J. Redondo, I. Fraga, I. Padrón, and M. Comesaña, "The Spanish adaptation of ANEW (Affective Norms for English Words)," *Journal of Behavior Research Methods*, vol. 39, no. 3, pp. 600–605, 2007.

[11] A. P. Soares, M. Comesaña, A. P. Pinheiro, A. Simões, and C. S. Frade, "The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese," *Journal of Behavior research methods*, vol. 44, no. 1, pp. 256–269, 2012.

[12] E. Palogiannidi, E. Iosif, P. Koutsakis, and A. Potamianos, "Affective lexicon creation for the Greek language." submitted to *SEM, 2015.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 168–177.

[14] K. Balog, G. Mishne, and M. De Rijke, "Why are they excited?: identifying and explaining spikes in blog mood levels," in *Proc. of Association for Computational Linguistics*, 2006, pp. 207–210.

[15] N. Malandrakis, A. Kazemzadeh, A. Potamianos, and S. S. Narayanan, "SAIL: A hybrid approach to sentiment analysis," in *Proc. of SemEval*, 2013, pp. 438–442.

[16] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "In the mood for being influential on twitter," in *Proc. of Privacy, Security, Risk and Trust (PASSAT) and IEEE Conference on Social Computing (SocialCom)*, 2011, pp. 307–314.

[17] F. Celli, "Unsupervised personality recognition for social network sites," in *Proc. of International Conference on Digital Society (ICDS)*, 2012, pp. 59–62.

[18] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[19] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proc. of Association for Computational Linguistics*, 2004.

[20] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using context to improve emotion detection in spoken dialog systems," in *Proc. of Interspeech*, 2005.

[21] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining." in *Proc. of LREC*, 2006, pp. 417–422.

[22] C. Strapparava and A. Valitutti, "WordNetAffect: an affective extension of WordNet." in *Proc. of LREC*, 2004, pp. 1083–1086.

[23] N. Malandrakis, A. Potamianos, E. Iosif, and S. S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2379–2392, 2013.

[24] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis." in *Proc. of Human Languages Technologies and Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[25] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[26] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[27] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition." in *Proc. of Interspeech*, 2002, pp. 873–876.

[28] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proc. of ACM*, 2012, pp. 485–492.

[29] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, 2011.

[30] A. Pargellis, E. Fosler-Lussier, C.-H. Lee, A. Potamianos, and A. Tsai, "Auto-induced semantic classes." *Speech Communication*, vol. 43, no. 3, pp. 183–203, 2004.

[31] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[32] D. Marquardt and R. Snee, "Ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[33] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[34] E. Iosif and A. Potamianos, "Unsupervised semantic similarity computation between terms using web documents." *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1637–1647, 2010.

[35] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[36] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: a computational study." *Behavior Research Methods*, vol. 39, no. 3, pp. 510–526, 2007.