Affective Words and the Company They Keep: Studying the Accuracy of Affective Word Lists in Determining Sentence and Word Valence in a Domain-Specific Corpus

Nadine Braun®, Martijn Goudbeek®, and Emiel Krahmer®

Abstract—In this article, we explore whether and how linguistic and pragmatic context can change individual word valence and emotionality in two parts. In the first part, we investigate whether sentence contexts retrieved from a domain-specific corpus (soccer) bias individual word affect. We then examine whether word valence with and without context accurately indicates sentence valence. In the second part, we compare word ratings from the first part to four different existing affective lexicons, with different levels of sensitivity to semantic and pragmatic context, and examine their accuracy in determining sentence valence. Results show a significant difference between words with and without context, the former more accurate in determining sentence valence than the latter. The preexisting lexicons were found to be similar to the individual word ratings collected in the first part of the study. With human-evaluated, context-sensitive lexicons being the most accurate in determining sentence valence. We discuss implications for emotion theory and bag-of-words approaches to sentiment analysis.

Index Terms—Affect analysis, affective word lists, bag-of-words methods, emotional corpora

1 INTRODUCTION

 LANGUAGE does not only serve the purpose of conveying objective information but often does much more than that. Words can also be an expression of a writer’s intentions and emotional experiences that resonate in their words, which, in turn, can shape the understanding of a reader. Often, these words do not carry an inherent affective meaning but are influenced by the company they keep [1] in that they are influenced by the pragmatic and semantic context, consisting of the words (frequently) surrounding them and the situation in which they are produced [2], [3]. This observation has implications for computational sentiment analysis, a line of research dedicated to automatically detecting the valence (positive or negative) and emotional stance of texts (for a discussion of the terminology see [4]), with a range of applications, from marketing and customer opinion detection to health monitoring and psychological status detection [5]. A common approach to sentiment analysis is to rely, either explicitly or implicitly, on a so-called “bag of words” assumption and treat individual words as separate units of affective meaning. By combining the scores of the individual words in a text, the affective meaning of the text can then be determined. For this purpose, a multitude of affective words lists have been developed. These word lists usually contain ratings of the words’ valence levels, sometimes also ratings for arousal, i.e., the degree of excitement, and dominance, i.e., the degree of control, based on the VAD dimensions ([6], [7]), or categorizations of words (or phrases) into emotion categories, which are often based on basic emotions [8], made by human judges. More modern variants of sentiment analysis often rely on advanced machine learning techniques (most notably neural networks). Based on annotated training material, they try to learn which words (or word-embeddings; a theme to which we will return below) are associated with which affective states, and then attempt to classify new texts accordingly (e.g., [9], [10]). In certain methods, bag-of-words and other representations are combined as inputs for machine learning methods, giving rise to hybrid approaches (e.g., [11], [12]). For example, Giatsoglu et al. [13] combined bag-of-word with word2vec representations [14]. Other studies explore the benefits of BERT-based models [15] for sentiment analysis. While such more complex methods may currently obtain state-of-the-art results on benchmark datasets [16], word-based approaches are still frequently used and have some important advantages over more complex, black-box models: they are easy to use and understand, and are transparent about how words are rated and to which affective categories they can be assigned. What these approaches have in common is that they start from the affective meaning of individual words. Affective lexicons are compilations of these words. Yet, it is unclear how accurately such dictionary approaches assess the affective meaning of texts.

2 THEORETIC FRAMEWORK

That the affective meaning of words can be influenced by the semantic context in which they appear is a notion...
reflected to in corpus and cognitive linguistics as “semantic prosody” ([2], [17]). The term describes the effect that affective meaning can be transferred from frequent collocations in which a word appears to the word itself, in that the negative or positive connotation can remain even though the word appears in combination with other words. Although the details of the effect are much debated (see, e.g., [18]), that some words have a positive or negative connotation is generally acknowledged. Take, for instance, the word “shoot”. In both the British National Corpus (BNC; [19]) and the Contemporary Corpus of American English (COCA; [20]), the top-ten-collocation lists are dominated by words like “down”, “kill”, “gun”, and “myself”. Consider the following constructed examples:

1) She shoots the man.
2) She shoots the ball.
3) She shoots the ball into the goal.
4) She shoots the ball into her own team’s goal.

Seeing a word like “shoot” in isolation is likely to illicit a negative association in the head of a reader, which would lead to a negative rating of the word if presented by itself. The negative connotation would probably be rooted in a use of the word “shoot” as in (1) where this particular word sense has the implicit association with fire arms and crimes as exemplified by the frequent collocations retrieved from the BNC and the COCA. However, the same word is also frequently used in much more neutral contexts, such as (2), or even more positively connotated ones, like (3). Yet, (3) can also be negatively connotated within the same domain (soccer), as shown in (4). Taking this thought even further, the perception of (3) as positive and (4) as negative might additionally depend on the reader’s affiliation with the shooter and her team. Especially example (4) illustrates that the different affective connotations for (1) – (3) are not merely associated with different senses of the word “shoot” or its collocations. Being able to detect such affective nuances in language is an important task during everyday interpersonal communication and, hence, should be taken into account in automatic systems aimed at detecting emotions, e.g., in text. This appears to be especially crucial when making inferences about the author’s emotional state, especially if these systems are used for diagnostic purposes ([21], [22], [23]).

The question is how bag-of-words approaches to sentiment analysis can fit into such a theoretical framework. If the affective meaning of words does not only depend on different word senses but also on connotations transferred onto them by frequent negative and positive collocations, and if speakers of a language are often not consciously aware of the latter effect while, at the same time, not being offered a way to determine the target word sense, prosodies and dominant word senses are likely also traceable in averaged word ratings comprising affective lexicons. Indeed, as examples (1)-(4) show, the averaged word rating collected from a number of raters indicates an underlying negative bias to the word, although more neutral highly frequent collocations (e.g., “Shoot a photo”) exist as well. In that sense, dictionaries for automatic sentiment detection in text are not intended to represent the only affective meanings of words and, thus, cannot and should not be considered “ground truth”. Much more, they can be regarded as an attempt at determining an a priori affect for each word – to capture interesting valance and emotion patterns in large amounts of text (e.g., a positivity bias, [24], [25]). In fact, in the context of laboratory studies, these averaged valences show consistent effects on participants: reaction times to negative words tend to be slower than reactions to neutral or positive words [26], eye-tracking studies show longer fixations on highly valenced than on neutral words [27], neural responses to affective words are stronger compared to neutral words ([28], [29]), and affective words are remembered better than neutral ones [30]. Furthermore, in a principal component analysis, the computational model proposed by Hollis and Westbury [31] identified a component related to the dimension of valence to organize lexical meaning, which can be interpreted in a similar way as averaged human valence ratings of individual words. Overall, bag-of-words approaches using averaged ratings of isolated words are able to reveal meaningful patterns in language. This is probably possible not because these approaches disregard the influence of the immediate semantic context of words. Rather, some context is already integrated in the word ratings in the form of specific word senses and semantic prosodies.

Although averaged word ratings offer a nice starting point for several lexical sentiment analysis techniques to investigate patterns of affective meaning in large corpora, attempts have been made to take the textual surroundings of words into account (see, e.g., [32]). Yet, despite a few notable exceptions, the influence of sentence context on the affective understanding of words has received little systematic attention in affective computing compared to other fields like cognitive linguistics, and has mostly been restricted to the effects of so-called contextual valence shifters [33]. These include negations, diminishers, and intensifiers [34], which are able to reverse, decrease, or increase the valence of single words and have to be considered for the correct interpretation of a word and its affective meaning. For example, while “good” in “a good idea” is undeniably a positive word, preceding it by “not” reverses the expression’s meaning and, hence, its perceived valence in the context. In addition, there are modifiers that can change the intensity of an emotion word, such as “very”, “rather”, or “slightly”. These specific phenomena are well-known, and various automatic sentiment analysis tools try to take them into account (e.g., [35]). Other features impacting valence (and emotionality more generally) have been less explored in prior research. On the phrase level, Wilson et al. [36] compared the contextual valence of multi-word expressions (categorically) to their polarity outside of a textual context, as well as their influence on the performance of a classifier algorithm. They observed a switch of polarity from negative and positive to neutral in many instances for expressions in context. This switch also impacted the performance of features of their machine learning algorithm negatively. Further, the combination of all textual features improved the accuracy of the polarity classifier the most in the study, suggesting that valence and emotionality indeed depend on textual factors beyond negations, diminishers, and intensifiers. However, since the study used expressions sorted into categories rather than valence ratings of individual
words, the question remains to which degree context biases the valence of individual words, which are predominant in dictionary approaches. Hence, the first connected research questions we are addressing in the current study are related to the quantification of in- and out-of-context differences in valence ratings of individual words:

RQ1: To what extent does the linguistic context of a sentence bias valence and emotion ratings of individual words? If valence changes as a function of context, are words rated in context more precise indicators of sentence valence than words rated in isolation?

Not all affective lexicons are constructed the same way or have the same aim. Some are categorized into valence or emotion categories; others are rated on scales for valence, arousal and dominance (VAD). Some of these dictionaries are further integrated in sentiment analysis tools. For example, the Linguistic Inquiry and Word Count (LIWC; [37]) is widely used in social sciences to investigate, among others, affective language use, signs of depression in language [38], and personality in language [39]. LIWC’s approach is simple: it counts words that belong to one or more predefined categories (about 90), ranging from pronouns and positive or negative emotion words to words belonging to specific concepts, such as biological processes (body, health, sexuality) or relativity (time, space). Since it has been first developed at the beginning of the 1990s, the original English dictionary has been repeatedly updated and translated to many different languages and is frequently used and cited. One of the positive aspects of LIWC and similar tools lies precisely in its dictionary: in contrast to some machine learning approaches, it is easy to use and transparent in how words are rated and which categories they belong to. These category entries have been extensively validated by human judges (see [40] for a detailed explanation of the seven validation steps taken for LIWC 2015), including the positive and negative affect categories as well as the discrete negative emotion word categories, which are anger, anxiety, and sadness.

In contrast to LIWC, other lexicons tend to be larger and often contain words rated on VAD scales. For example, the lexicon created by Warriner and Kuperman [41] contains almost 14,000 English words that have been rated on valence, arousal, and dominance by human judges. The Warriner lexicon is an updated and expanded version of the Affective Norms for English Words (ANEW) by Bradley and Lang [42], which has been used in a variety of linguistic affect studies (e.g., [43], [44]). Differently from LIWC, it is not integrated in an existing sentiment analysis tool but its entries are usually used as experimental stimuli or training material for computational models. Using a similar approach, a rather new affective lexicon created by Mohammad [45] comprises about 20,000 entries also organized on the VAD dimensions. However, the entries were not rated on metric Likert scales but ordered on best-worst scales for each dimension, yielding ordinal scores for each word. Since it takes semantic context into account to some degree, this approach intends to allow for a better representation of the affective nuances of words. While the three lexicons described above are intended as general purpose and domain-independent, there also exist domain-sensitive affective dictionaries. For example, Hamilton et al. [46] created word lists based on topics of Reddit, a popular online forum with many sub-communities. The lexicon contains affective word lists for the 250 biggest communities on the platform. The authors argue that taking the domain of the texts to be analyzed into account, as in the topic of the discourse such as sports or politics, is crucial for accurate results.

Considering the multitude of affective word lists, some more context-sensitive than others, we expect the accuracy, i.e., the lack of bias compared to the gold standard, and precision, i.e., the absence of noise, of different lexicons in determining text sentiment to be impacted to different degrees by domain-specific contexts. Therefore, the following related research questions guide the second part of this study:

RQ2: To what extent do ratings obtained for the current study match existing contemporary affective word lists? Do lexicons that are more sensitive to pragmatic and semantic context determine sentence valence more precisely than general purpose lexicons?

The current study aims at quantifying differences between the affective meaning of words with and without sentence context in a domain-specific corpus (soccer). The domain-specificity of the corpus is thereby intended to address the issue of extralinguistic context, which can pose a challenge for dictionary approaches in addition to linguistic context. We compare existing affective lexicons and newly collected word valence ratings with and without sentence context and explore their relation to the overall valence of the sentence they appear in. While the notion that context can have a large impact on the meaning of words has a long tradition (dating back at least to Wittgenstein [47]), few attempts have been made at quantifying contextual changes in the affective meaning of words, although potential biases can impact the accuracy of bag-of-words approaches to sentiment detection in text. This paper aims to tackle this.

3 PART 1: WORD RATINGS WITH AND WITHOUT SENTENCE CONTEXT

3.1 Design and Materials

To investigate the influence of context on word valence and emotionality, we let human judges rate a list of frequent words derived from a text corpus of a specific genre that was assumed to naturally contain high and low valence words and sentences. We focus on a corpus of sports reports [48], which consists of original soccer match reports in English, German, and Dutch and the matching game statistics, all published shortly after the games and taken from the involved football clubs’ websites. The English corpus, which was used for the current study, contains about two million words from 2,950 texts, which can be divided into reports on won (1,093 texts, 794,562 words), lost (1,091 texts, 733,573 words), and tied (766 texts, 529,259 words) matches. Exploratory analyses (see Table A1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2020.3005613) of the corpus with the affective lexicon by Warriner et al. [41] indicated that texts from the win subcorpus compared

1. The corpus is available for research purposes on request: https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/H00f3U
to aligned loss texts (i.e., texts written from the perspective of the respective losing teams) were generally lower on valence, with ties falling in between both categories, suggesting that “win” and “loss” can serve as operationalizations of positive and negative affective states, respectively. Therefore, in this paper, we focus on words and sentences taken from the subcorpus which are most likely to contain affectively meaningful words: the win and loss subcorpus of MEemoFC [48].

For the rating task, we chose keywords instead of simple frequency to find the most frequent words that are indicative of the respective subcorpus and, hence, likely to be affect-laden. This way, common stop words, such as articles or modal verbs were filtered out automatically. Keywords were determined by comparing word frequencies based on word totals in both subcorpora using log-likelihood ratios (Antconc; [49]). From the keyword lists, the first 250 entries were chosen, respectively, and controlled for undesirable/inconvenient words, such as pronouns, prepositions, (nick) names, abbreviations, acronyms and abbreviations, numbers, question words, and deictic terms (e.g., “this”, “here”, or “today”), as these words are unlikely to be valenced or affective without context. For every word removed from the initial 250-word win/loss lists, we moved up in the keyword lists and added the earliest substitute. The resulting list was controlled for prevalence, i.e., how well a word is generally known by people, of each word. We only selected words that were known by 70 percent or more participants in the study of Brysbaert et al. [50] because the knowledge of a word is a logical prerequisite for rating its affective potential. In the end, all words had a minimum familiarity of 80 percent, with the vast majority higher than 95 percent. Other than the mentioned requirements, there were no exclusion criteria and no focus on specific word types.

Since our study is not only concerned with isolated words, the occurrences in the whole corpus for each word were searched for in a concordance tool [49] and, out of all sentences, two were randomly chosen as contexts. We then excluded ungrammatical sentences, headlines, or sentences that already contained valence modifiers. Each excluded sentence was substituted with another randomly chosen one. These sentences were then controlled for the exclusion criteria. During the context selection, additional words had to be excluded because they either turned out to be (part of) player or place names, or due to only one appropriate sentence that matched our criteria being available. The same sentence contexts could be used for different words. In total, 11 out of 1000 sentences were duplicates. Due to a procedural error, one word appeared in the word list twice. The duplicate and its contexts were excluded post-hoc. All results are based on 499 words (250 from the loss subcorpus of MEemoFC, 249 from win) and 998 sentence contexts.

3.2 Data Collection

The data was collected with surveys in Qualtrics (Provo, UT) and participants were recruited on Amazon’s Mechanical Turk. At the beginning of the survey, MTurk workers were informed about the aim of the study and asked for consent. If they proceeded, they first had to indicate their current affective state on four nine-point scales that controlled for valence and arousal (happy – unhappy, satisfied – unsatisfied, wideawake – sleepy, excited – calm) to exclude a possible influence of the participant’s mood on their ratings. In the first half of the survey, every respondent rated 40 words (20 from the win and loss subcorpora, respectively) on valence scales, ranging from 1 to 9, and on 9-point scales for the three discrete emotions (anger, anxiety, sadness) that are also included in the LIWC [37]. Participants were instructed as follows: “Please indicate on the scales below in how far the following word without context, just on its own, is positive or negative/expresses the following emotions.” In the second part of the survey, they rated the same words in sentence contexts on the same scales (“Please indicate on the scale below in how far the highlighted word is positive or negative in the context of the sentence/expresses the following emotions.”). Additionally, they indicated the overall valence of the sentences on 9-point scales. In total, each participant rated 80 items. In both parts of the survey, items were randomized, and participants were not informed that they would rate the same words in a soccer context later on. Additionally, two attention checks were included in the context rating part to ensure that participants actually read the sentences they rated. These checks had the same form as the critical sentence contexts but instead asked respondents to choose the middle value on all scales. The survey concluded with questions about demographics (age, gender, education, the participant’s native language(s) and a nine-point scale to indicate whether they enjoyed watching or playing soccer). Participants were paid $3.00 for the completion of the survey. To cover all 499 words and the respective sentence contexts, we created 25 survey versions.

As, due to the setup of the surveys, overall sentence valence ratings could be influenced by individual word ratings, we created an additional survey to obtain independent valence ratings for the sentences. The beginning and end of the surveys were exactly the same as for the combined surveys, but respondents only rated 45 sentences including two attention checks on 9-point valence scales. The items were randomized, and MTurkers were payed $2.50 for completion.

3.3 Participants

Access to the task on MTurk was limited: workers 1) had to be located in English-speaking countries; 2) had to be native-level English speakers; 3) had been granted the Amazon Mechanical Turk Masters qualification, which is used by Amazon to distinguish workers who generally perform well on the platform; 4) had at least a 97 percent task approval rate (initially 95 percent; raised due to workers failing the attention checks) for previous human intelligence tasks (HITs). After a survey was finished, the participating workers were excluded from the next survey batch. Five workers were assigned to each individual survey.

In total, 125 workers completed the 25 combined surveys (66 male, 58 female, 1 other), $\mathrm{Age}_{\bar{M}} = 39.00$ years ($SD = 11.03$). The educational background of worker was mixed: 17 indicated high school education, 36 some college, 63 a Bachelor’s degree, 6 a Master’s degree, and 3 had a Doctorate degree. All were native English speakers, with only 6 mentioning a second native language (2 Korean, 4 Spanish). On average, participants reported feeling rather happy ($M$...
Table 1: Comparison of Individual (IR) and Context Ratings (CR) by Affective Categories (Valence, Anger, Sadness, and Anxiety)

| Subcorpus | Mean (IR) | SD (IR) | Mean (CR) | SD (CR) | b | 99% CI | SE |
|-----------|-----------|---------|-----------|---------|---|--------|----|
| Valence   |           |         |           |         |   |        |    |
| Win       | 6.30      | 2.02    | 6.43      | 2.02    | 0.13 | 0.05, 0.04 | 0.22 |
| Loss      | 4.28      | 2.28    | 4.36      | 2.22    | 0.08 | 0.09, 0.04 | 0.16 |
| Sadness   |           |         |           |         |   |        |    |
| Win       | 1.49      | 1.29    | 1.49      | 1.33    | -0.004 | -0.15, 0.04 | 0.02 |
| Loss      | 2.70      | 2.46    | 2.64      | 2.37    | -0.07 | -0.37, 0.05 | 0.19 |
| Anxiety   |           |         |           |         |   |        |    |
| Win       | 2.06      | 1.95    | 2.08      | 1.95    | 0.02 | -0.07, 0.04 | 0.09 |
| Loss      | 3.17      | 2.54    | 2.89      | 2.32    | -0.28 | -0.37, 0.05 | 0.05 |
| Anger     |           |         |           |         |   |        |    |
| Win       | 1.48      | 1.27    | 1.44      | 1.21    | -0.04 | -0.13, 0.04 | 0.04 |
| Loss      | 2.29      | 2.12    | 2.16      | 1.92    | -0.13 | -0.20, 0.04 | 0.04 |

Note: Significant comparisons in bold.

3.4 Results

As expected and intended, individual words (M = 6.30, SD = 2.02) and sentences (M = 6.46, SD = 2.02) taken from the win subcorpus of MEmoFC [44] were rated higher on valence than words (M = 4.28, SD = 2.28) and sentences (M = 4.41, SD = 2.17) from the loss subcorpus (total for words: M = 5.29, SD = 2.38; total for sentences: M = 5.43, SD = 2.33). All ratings were normally distributed. Words from loss were also rated higher on anger (M = 2.29, SD = 2.12), anxiety M = 3.17, SD = 2.54, and sadness M = 2.70, SD = 2.46) than in win (anger: M = 1.48, SD = 1.27; anxiety: M = 2.06, SD = 1.90; sadness: M = 1.49, SD = 1.29).

A similar picture emerges for the independent sentence ratings: sentences from the win subcorpus were again rated higher on valence, M = 6.43, SD = 1.58, than sentences from the loss subcorpus, M = 4.65, SD = 1.71, with the overall mean being close to the middle of the scale, M = 5.54, SD = 1.87.

To investigate the potential difference in ratings for individual words and words in context, we ran linear mixed effects models in R using the lme4 package [51] to account for word and participant variation. Words and participants were included as random intercepts in the model with the interaction of context (yes/no) and subcorpus (win/loss) as the predictor and word valence (ranging from 1 to 9) and the discrete emotions (anger, anxiety, and sadness; ranging from 1 to 9) as the dependent variables, respectively. As models including random slopes for participants and items failed to converge, likely due to the low number of observations, slopes were removed subsequently. To assess variation in the standard deviation of the ratings, the model included the type of rating (individual/in context) and subcorpus as predictors and words as random intercepts. All final models were bootstrapped over 100 iterations to yield confidence intervals and standard errors.

Results show that words without context are rated significantly lower on valence compared to words with context (Table 1). The interaction between subcorpus and type of rating was not significant. This small but significant change in valence due to sentence context hints at a context-related bias, which potentially affects the precision of the words in determining sentence valence. However, it should be noted that there is much variability in the data, depending on each the word and its contexts: words positive in isolation can be rated to be more negative in one sentence context, while they are perceived as more positive in another. A similar pattern emerges for words negative in isolation. These different trends are illustrated in Fig. 1 for both subcorpora. In the loss subcorpus, the trend is mixed: many words in context are rated higher and lower in valence than by themselves, and fewer words remain at the same valence. In the win contexts, the pattern is different: more words in context are rated higher in valence or remain at the same valence level as in isolation, while fewer words are rated lower on valence than in loss. The relation between individual word valence and words in context is also depicted in Figs. 1 and 2. Likewise, the respective Spearman correlation of the two
averaged measures, $r_S = .76, N = 499$, is lower than what would have been expected (i.e., a correlation of .90 or even higher) if sentence context had no influence on the individual word ratings. In addition, as illustrated by the larger standard deviation for context ratings, agreement of participants for the valence ratings is significantly lower for words in context ($M_{SD} = 1.49, SD = 0.51$) compared to individual word valence ratings ($M_{SD} = 1.13, SD = 0.45, b = 0.36, SE = 0.03, BC 95% CI [0.30, 0.41]$). This effect did neither differ by subcorpus nor was there a significant interaction with type of rating, which makes it more difficult to determine the direction of valence shifts for both subcorpora with certainty.

The ratings of discrete emotions appear to depend on context as well, although the pattern is different than the one for valence. For words rated without context on anger, the difference with words in context is only significant for loss, showing that anger decreased when assessed in the context of sentences (Table 1). The interaction between subcorpus and type of rating was not significant for anger. For anxiety, the pattern is similar: words without context scored significantly differently in context only for loss, not for win, showing that anxiety ratings decreased in the context of sentences (Table 1). Here, subcorpus and type of rating interacted significantly ($b = 0.30, SE = 0.06, BC 95% CI [0.18, 0.43]$). The interaction was primarily driven by slightly higher anxiety ratings in context word ratings in win ($M = 2.08, SD = 1.95$) compared to individual word ratings ($M = 2.06, SD = 1.90$), in contrast to a decrease of anxiety ratings of individual words in loss ($M = 3.17, SD = 2.54$) compared to context word ratings ($M = 2.89, SD = 2.32$). In contrast to anger and anxiety, the differences between sadness ratings of words without context and words with context were not significant, showing that the sadness ratings remained approximately the same independent of context (Table 1). Consistent with the increase in valence ratings, anger and anxiety of words also decreased significantly in our chosen sentence contexts. Therefore, context even within one domain seems to be able to bias the perceived valence and emotionality of individual words. Yet, the question arises how both types of ratings relate to the valence of the sentences the words appear in. To investigate this, we used Spearman’s correlations over the averaged ratings per word and sentence from the combined surveys. In addition, we correlate the ratings with the independent sentence ratings to exclude the possibility of a carry-over effect from word valence to sentence valence in the first surveys.

Overall, while both types of ratings correlate strongly with sentence valence, the higher correlation shows that ratings in context ($r_S = .92, N = 998$) seem to be much more accurate than isolated ratings ($r = .59, N = 998$). A similar effect emerges for the independent sentence ratings and words with ($r_S = .72, N = 998$) and without context ($r_S = .46, N = 998$). This is illustrated in the respective heatmaps (Fig. 3). While the trend similar in all heatmaps, although more accurate in context ratings, it is more streamlined and, thus, also more precise for words in context. This suggests that a bias which could not be detected by the isolated word ratings was introduced by the sentence contexts. Moreover, the patterns remain largely the same for both subcorpora, hinting at an overall higher precision of words in context in determining sentence valence, which is not only driven by part of the data.

### 3.5 Discussion

In the first part of this study, we used a domain-specific (soccer) keyword list, representing pragmatic context, and showed that the presence of a sentence, representing linguistic context, significantly biased valence ratings and ratings for individual emotions (anger and anxiety) of individual words and that these changes could even be differentiated by the subcorpora. Since we selected our sentence stimuli from a specific domain, and in such a way that no obvious valence shifters, diminishers, or intensifiers could influence the word valence in the sentence, it is likely that not only the semantic context of the words caused changes in valence and emotionality, but also their pragmatic context. This implies that, although the general direction of valence determined by bag-of-words approaches might be accurate, their accuracy and precision in sentiment analyses might be lower compared to contextualized affective lexicons.

Since lexicons with different kinds of valence measures exist (categorical, scalar, ordinal), which consider context to different degrees already, we chose lexicons that we expected to be gradually more context-sensitive in the next section. In addition to more semantic sensitivity, we include a dictionary that focuses on sports for our specific domain “soccer”, which might also offer a better approach to possible confounding factors such as homonymy or polysemy.

### 4 PART 2: ACCURACY OF EXISTING AFFECTIVE LEXICONS

#### 4.1 Design and Materials

To investigate to what extent the word ratings obtained in the first part match contemporary affective lexicons and how strongly the entries in these lexicons are related to the valence of our selected sentences, we chose four different affective lexicons that we expect to be context-sensitive in different ways and to different degrees: 1) the affective categories of the most recent English LIWC dictionary [37] (affective, positive emotions, and negative emotions), 2) the
word list created by Warriner et al. [41] (in the following referred to as Warriner Lexicon, WL), 3) the ordinal lexicon by Mohammad [45] (in the following referred to as the NRC-VAD), and 4) a wordlist extracted from the Reddit sub-community for sports by Hamilton et al. [46] (in the following referred to as r-sports; see Table 2). While, to the best of our knowledge, LIWC and the r-sports were developed independently from each other and the other lexicons, the NRC-VAD also incorporates terms from the WL. There are a number of important differences between the dictionaries. The WL consists of a total of 13,915 words, for which online ratings have been crowdsourced on 9-point scales for the VAD dimensions. The affective part of LIWC comprises a total of 1,416 words, with 641 classified as positive, 745 as negative, an additional 29 words only as affective without further categorization. It should be noted that LIWC also contains open or wildcard words that are marked with an asterisk, such as, e.g., *ador* or *glory*, which encompass all possible word endings for the respective morphemes. The NRC-VAD was compiled with best-worst scaling and includes entries from a range of earlier lexicons. The r-sports is a learned lexicon, compiled by automatically extracting affective words from the Reddit sport sub-community.

4.2 Analysis

We compared the domain-specific word list from Part 1 to the entries in the LIWC, the WL, the NRC-VAD, and the r-sports. For the words that overlap with the lists, we examined whether our isolated word ratings matched the ratings and categories assigned to them in the respective lexicons with Spearman correlations. Finally, we assessed the ability of the four lexicons in determining sentence valence.

Fig. 3. Heatmaps indicating the frequencies of average individual (a, b) and word-in-context (c, d) valence ratings relative to average sentence valence from the combined (a, c) and independent (b, d) sentence ratings.

2. http://crr.ugent.be/archives/1003
3. https://salfmohammad.com/WebPages/nrc-vad.html
4. https://nlp.stanford.edu/projects/socialsent/
5. Raters were presented with four words (4-tuples) at the same time and had to order these words from highest to lowest valence (arousal and dominance). Then the proportion of number of times a word was selected as the lowest on valence/arousal/dominance was subtracted from the proportion of times it was chosen as the highest on the dimensions. These scores were then linearly transformed to a 0 to 1 interval.
compared to our words-in-context ratings with sentiment analyses for each list per sentence, taking into account all words in each sentence.

4.3 Results

The number of words from our word list that overlaps with each lexicon differs greatly: out of 499 words, LIWC contains 114, the WL 437, the NRC-VAD 460 words, and the r-sports 238 (Table 2). To correlate the categorical LIWC with our word scores in a similar way as the other three lexicons, we assigned numbers to the LIWC categories, following the 9-point Likert scales we used in Part 1: positive = 9, negative = 1, and affective = 5. Of course, this implies “extreme” numeric ratings for all words in LIWC, although, for example, not all words in the positive emotion category are likely to be equally positive. The lexicon ratings correlate more strongly with the individual ratings (LIWC : \( r_S = .87 \); WL : \( r_S = .89 \); NRC-VAD : \( r_S = .89 \); r-sports : \( r_S = .24 \)) obtained in our study than with the context ratings (LIWC : \( r_S = .83 \); WL : \( r_S = .64 \); NRC-VAD : \( r_S = .68 \); r-sports : \( r_S = .20 \)). Hence, the words that were presented to our participants individually were rated largely the same as they were rated and categorized by the participants that evaluated the entries of the LIWC, WL, and NRC-VAD. To a lesser extent they also match the learned valence scores in the r-sports.

Since the existing ratings in all four lexicons are more similar to our individual word ratings, their relation to the overall sentence valences is likely similar as well. To compare the lexicons’ accuracy in determining sentence valence, with each other but also with our words in context, we ran sentiment analyses by first tokenizing the sentences and then searching for matches in each list for each sentence. We then calculated a mean valence score for all matches per sentence. For LIWC, we ran the analysis for positive and negative emotions in the original program for each sentence, which returned proportions of positive and negative emotion words per sentence, respectively. For the returned scores, we subtracted the negative emotion scores from the positive ones to yield overall sentence valence scores. We then used a Spearman correlation to compare all lexicon scores to our human independent sentence ratings. The relation between word list scores and human sentence ratings is illustrated in Fig. 4 (scales turned into z-scores for comparability). All word lists, surprisingly with the exception of the r-sports, show strong correlations with sentence valence, which indicates that, generally, accuracy was acceptable and comparable for the LIWC, WL, and NRC-VAD. Further, the NRC-VAD, which was expected to be the most sensitive to semantic context, and our in-context word ratings, which were sensitive to both, semantic and the domain-specific context (soccer), are the most strongly correlated approaches, indicating higher accuracy in sentiment detection.

4.4 Discussion

In Part 2, we showed that the isolated word ratings collected in Part 1 largely matched the ratings and categorizations in

| Lexicon   | Entries | Type of rating | Overlap with our word list (out of 499 words) |
|-----------|---------|----------------|---------------------------------------------|
| LIWC (affective) | 1415 | Categorized (positive, negative, affective, hierarchical) | 114 |
| WL        | 13,915 | VAD (scale, human) | 437 |
| NRC-VAD   | 20,007 | VAD (ordinal, human) | 450 |
| r-sports  | 4,571  | Valence (automatically extracted with SentProp) | 238 |

Fig. 4. Relationships between independent sentence valence ratings (z-scores) and z-scores of valence analyses returned per sentence by the lexicons: a) LIWC (total), b) WL, c) NRC-VAD, d) r-sports, and e) words in context.

TABLE 2
Properties of the LIWC, WL, NRC-VAD, and R-Sports
four existing contemporary affective lexicons, although neither of the four dictionaries contained all of our keywords. The sentiment analysis done on the sentence level with the WL, the LIWC, and the NRC-VAD returned scores with accuracies that highly correlated with the independent human sentence ratings and were increasingly precise in determining sentence valence. Although the r-sports lexicon was supposed to be the most appropriate lexicon for our corpus because it was extracted from a sub-community of Reddit dedicated to sports, it yielded the lowest correlation scores. This could be explained by the fact that the r-sports was the only learned lexicon and was not evaluated by human judges. Therefore, despite the impressive progress achieved in computational sentiment analysis, for the construction of affective lexicons, human ratings still appear preferable. Even considering all words in each sentence, the context word ratings obtained for Part 1 were better indicators of sentence valence than the much longer existing lexicons.

Thus, the precision of affective lexicons can benefit from taking their intended domain and use into account during construction. A change as a function of context, pragmatic or semantic, can lead to a mismatch of the rating or category assigned to a word in isolation and its assessment in an actual text. This becomes obvious in the following examples, which are taken from the LIWC analysis but should be taken as representative of challenges all bag-of-words approaches face. The underlined words are the ones included in the affective LIWC, with words in italics being the focus words for the rating task of human judges:

1) TEAM1 fought off a TEAM2 fightback to claim a thrilling victory at PLACE.
2) Heartbreak hit in the 92nd minute though when PERSON’s late winner secured the spoils for the TEAM.
3) PERSON’s goal in the early stages of the second half wrapped things up for the TEAM1, who bounced back from defeat to TEAM2 on the opening day.

In (5), the focus of human raters in our surveys was the word “thrilling”, to which they assigned a valence of $M = 7.8$ (SD = 1.13; individually) and $M = 8.6$ (SD = 0.55; in context), while the overall sentence was rated to be an $M = 8.6$ (SD = 1.73) on average. However, by design, LIWC only recognized “thrilling” as positive and categorized “fought” and “fightback” as negative emotion words in the example sentence, which distorts the affective intention of the sentence towards negative, effectively lowering the tools accuracy. Similarly, raters focused on “heartbreak” in (6), which was assigned $M = 1.3$ (SD = 0.48) in isolation, remaining at $M = 1.2$ (SD = 0.44) in the context of the sentence, with the overall sentence being rated at $M = 2.4$ (SD = 1.14). Yet, LIWC recognizes three words out of which two are positively categorized (“win” and “secure”), which results in a higher proportion of positive emotion words. For (7), the problem is different one again: while LIWC only classifies “defeat” as negative, and therefore returns a bigger proportion of negative emotion words for this sentence, participants, who also focused on “defeat” rated the word in isolation, in accordance with LIWC, at $M = 1.8$ (SD = 1.32), but $M = 6.2$ (SD = 1.92) in the sentence, which was rated to be positive with $M = 7.4$ (SD = 1.67). In this example, carry-over seems to have happened from the sentence valence to the word valence or the complete expression “bounce back from defeat”, which serves a similar function as a regular negation. These contextual changes likely not only affect the words in our word list but also other entries in the lexicons.

5 GENERAL DISCUSSION AND CONCLUSION

In the present study we set out to quantify the impact of sentence context on the affective rating of individual words. In Part 1, we investigated to what extent the context of a sentence changes valence and emotionality ratings of individual words and how well word valence, individual and as a function of sentence context, relates to the valence of sentences the words appear in. To do so, we used a domain-specific corpus, the MEMoFC [48] for soccer reports. We found a significant difference induced by sentence context in that many words were rated drastically different in isolation. Further, words rated in context were more strongly correlated with the valence of the sentences they occurred in. Interestingly, agreement between participants was lower for words rated in context compared to words rated in isolation, which nicely illustrates that valence not only depends on the semantic context of a word but also on its pragmatics, e.g., on the reader, their world knowledge, and additional subtle nuances resonating in a sentence. By collecting separate sentence ratings without highlighting individual words, we also showed that a simple carry-over effect from the word to the sentence level was not enough to explain the stronger association of words in context with sentence valence.

It is likely that not just the complete sentence changes the affective meaning of individual words but already multi-word constructions (n-grams) can shift word valence and emotionality. Computationally, some sentiment analysis approaches (e.g., [5], [32]) do indeed take n-grams into account, but, to our knowledge, differences in affect between individual words and multi-word expressions have not been investigated systematically yet. Exploring potential changes of affective meaning from isolated words to multi-word expressions to full sentences in different domains might be an interesting starting point for future research and the construction of future lexicons. While we investigated contextual valence changes in individual words with respect to sentences, their influence on longer texts might be different and should also be addressed in future research. Considering that studies have shown that unigrams, i.e., single words, are more accurate predictors of valence in short texts like microblogs while longer texts such as full blog entries benefit from taking longer expressions into account in analysis [49], contextual changes might be able explain such results.

In Part 2, we examined to what extent ratings obtained for the current study matched four contemporary affective lexicons, each gradually more context-sensitive: LIWC [37], a lexicon by Warriner et al. [41], a lexicon containing ordinal ratings [45], and a domain-specific lexicon extracted from a subcommunity of Reddit dedicated to sports [46], which we assumed to be suitable for our soccer corpus. We also ran sentiment analyses on the individual sentences using all four lexicons and our context word list, and correlated the
returned scores for each sentence with our human “gold standard”. Results showed that, while all lists but the r-sports correlated strongly with sentence valence, there were also differences in accuracy. The correlation with sentence valence was slightly weaker for the WL compared to the LIWC, for which it was, in turn, weaker than for the NRC-VAD. Our words-in-context ratings were overall most accurate in determining sentence valence. These results imply that not only the number of entries in an affective lexicon impacts its precision, but also to what extent it takes semantic (NRC-VAD) and pragmatic (our domain-specific list) context into account. Finally, we also presented individual examples of how and why analyses can go “wrong” with a focus on contextual issues.

Overall, we demonstrated that the context of a word can have a significant impact on its perceived valence and emotionality, introducing insights from the field of semantic prosody and evaluation into the area of sentiment analysis and natural language processing (NLP). We found that the majority of word valence and emotion ratings became more positive when presented in a sentence taken from our soccer corpus, although the word ratings contained much variation. These changes are consistent with the results of Wilson et al. [36]. By directly comparing in- and out-of-context ratings of words, we managed to quantify these contextual changes. We argue that these changes can affect the accuracy of word-based approaches to sentiment analysis by focusing on one specific domain, “soccer”. Still, it seems valid to assume that changes of affective word meanings also occur in other domains and corpora. These patterns could be accounted for to some extent by creating more context-sensitive lexicons. In future research, this could be investigated systematically on a larger scale: can the context word ratings such as the ones obtained for soccer reports in the current study also be used to predict sentence and text valence more accurately than individual word ratings in other soccer corpora, perhaps even in reporting for different sports? For audio-visual data, different kinds of context are increasingly considered to improve the accuracy of sentiment analysis approaches. For the same reason, we recommend also taking contextual changes of affective word meanings—be they pragmatic or semantic—into account when using isolated word ratings as indicators of affect in text.

As mentioned in the Introduction, many current NLP applications for sentiment analysis depart from initialized models with pre-trained word representations such as word2vec [14] or GloVe [53], or with pre-trained language models like ELMo [54] or BERT [15]. Currently, BERT-like models reach state-of-the-art performances, typically by finetuning the pre-trained models to the task at hand [16]. What is especially interesting about these BERT models for our current purposes is that they are contextualized: every input word is represented by a vector which depends on the particular context of the word occurrence ([15], [55]). This makes these models prima facie interesting for the study of affective words and the company they keep. While we feel this is indeed an interesting line for future research, there are some complications. First of all, while it is quickly becoming clear that BERT and its ilk work remarkably well, it is not so clear whether these models work [55]. The models have no clear cognitive motivation, and they are so large that they become difficult to understand and work with, which explains why so many studies try to get a better understanding of the underlying reasons for the models’ performance [55]. As long as these models are not transparent, we conjecture that word lists will still be frequently used, certainly for ‘sensitive’ applications, like trying to predict whether someone is depressed [23] or possibly even suicidal [21], based on their word use. This is not to say that context information is not important for such affective lexicons; in fact, creators of such affective lexicons have long been aware of the limitations of looking only at isolated words, and our results also suggest that context-sensitive lexicons are relevant when trying to determine, say, the valence of a sentence. Thus, to take this semantic context into account in a transparent way, words can be compared directly to each other (as done with best-worst scaling [45]; for contextualized lexicons see also [56]). Further, to incorporate pragmatic context, ratings can be collected within the domain of interest or by specific raters. While acknowledging every kind of context is certainly an impossible task, tailoring affective lexicons to the research question at hand should be possible. Research aims, just like corpora, provide ways to limit the number of relevant word contexts. Although contextualizing lexicons is likely more cost- and time-consuming than using general-purpose ratings, this option should be considered for sensitive applications of sentiment analysis, but also if a lexicon is built with a specific domain in mind. Once a number of contextualized lexicons exists for specific purposes, these lexicons can be reused and maintained.

As the current study has shown, the interpretation of individual words and their valence and emotionality relies greatly on the company they keep, both in terms of other words and extralinguistic information surrounding them.

ACKNOWLEDGMENTS

The authors received the financial support for this work from The Netherlands Organization for Scientific Research (NWO), via Grant PR-14-87, which is gratefully acknowledged. We also thank Lorenzo Gatti for his help with the analyses.

REFERENCES

[1] J.R. Firth, A Synopsis of Linguistic Theory, 1930–1955, Studies in Linguistic Anal., Oxford: Basil Blackwell, pp. 1–30, 1957.
[2] B. Louw, “Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies,” Text and Technol.: In honour of John Sinclair, M. Baker, G. Francis, and E. Tognini-Bonelli, Eds, Amsterdam/Philadelphia: John Benjamins, pp. 157–176, 1993.
[3] S. Hunston and G. Thompson, in Evaluation in Text: Authorial Stance and the Construction of Discourse: Authorial Stance and the Construction of Discourse, London, UK: Oxford Univ. Press, 2000.
[4] M. D. Munezero, C. S Montero, E. Sutinen, and J. Pajunen, “Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text,” IEEE Trans. Affective Comput., vol. 5, no. 2, pp. 101–111, Second Quarter 2014.
[5] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proc. 42nd ACL, 2004, pp. 271–278.
[6] C. E. Osgood, “Dimensionality of the semantic space for communication via facial expressions,” Scand. J. Pers. Soc. Psychol., vol. 39, no. 6, 1980, Art. no. 1161.
[7] J. A. Russell, “A circumplex model of affect,” J. Pers. Soc. Psychol., vol. 39, no. 3, 1980, pp. 550–553 1992.
[8] P. Ekman, “Are there basic emotions?,” Psychol. Rev., vol. 99, no. 3, pp. 144–149, 1992.
[9] A. L. Maas, E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Poits, “Learning word vectors for sentiment analysis,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol., 2011, pp. 142–150.
Martijn Goudbeek received the PhD degree from the Radboud University Nijmegen, in 2007 for work done at the MPI for Psycholinguistics on the acquisition of auditory categories. He spent two years at the Swiss Center for Affective Science studying the nonverbal expression of emotion, with special interest in the vocal channel. Since 2008 he works at Tilburg University at the department of Communication and Cognition. His research interests include language production and perception, emotion, statistics, nonverbal communication, and interpersonal communication.

Emiel Krahmer received the PhD degree from Tilburg University, in 1995. He is currently a full professor with the Tilburg School of Humanities and Digital Sciences. In his research, he studies how people communicate with each other, and how computers can be taught to do the same, to improve communication between humans and machines.