Synesketch: An Open Source Library for Sentence-Based Emotion Recognition

Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic

Abstract—Online human textual interaction often carries important emotional meanings inaccessible to computers. We propose an approach to textual emotion recognition in the context of computer-mediated communication. The proposed recognition approach works at the sentence level and uses the standard Ekman emotion classification. It is grounded in a refined keyword-spotting method that employs: a WordNet-based word lexicon, a lexicon of emoticons, common abbreviations and colloquialisms, and a set of heuristic rules. The approach is implemented through the Synesketch software system. Synesketch is published as a free, open source software library. Several Synesketch-based applications presented in the paper, such as the the emotional visual chat, stress the practical value of the approach. Finally, the evaluation of the proposed emotion recognition algorithm shows high accuracy and promising results for future research and applications.

Index Terms—Natural language processing, text analysis, sentiment analysis, emotion in human-computer interaction

1 Introduction

The burst of new computer-based media for communication and expression has caused an increasing need for human-computer interaction to acknowledge human emotions. The work presented in this paper focuses on emotion recognition in English text. The fact that online communication is still to a great extent text-based justifies such an approach. Our domain is online communication in the form of short messages, such as real-time instant messages (chat), comments on social media and social networking sites, or microposts (e.g., tweets). Previous research (e.g., [1]) has shown that such communication is beneficial for social relations, and that users consider chat a medium for fulfilling, versatile, and engaging communication.

The proposed recognition algorithm classifies the text of a sentence according to the following emotional categories: happiness, sadness, anger, fear, disgust, and surprise [2]. The proposed algorithm estimates emotional weights for each emotional category (how intense the emotion is) in the form of a numerical vector. The vector is used to determine the dominant emotional type (the emotional type with the highest weight) and the overall emotional valence of a sentence (is the emotion positive, negative, or neutral). If the vector is zero or close to zero, we consider the sentence emotionally neutral. Users of our software may determine their criteria for neutrality.

To recognize emotions in sentences, we use a hybrid of a keyword-spotting method and a rule-based method. Our keyword-spotting approach is based on the use of a lexicon of words and expressions related to emotions. We consider that our main contribution is threefold. First, in order to construct a word lexicon, we use both the power of human judgment and the power of WordNet, a lexical database for English language [3]. Specifically, we employ a survey-based word lexicon to automatically search WordNet for all semantic relatives of the initial word set. Second, we take into account ":)"s, ">:O"s, and "ROFL"s through an extensive emoticon lexicon. Third, we try to overcome some of the problems associated with keyword-spotting techniques (see Section 2 for details) with several heuristic rules. We argue that the proposed technique is suitable for analyzing fragmented online textual interaction that is abundant in colloquialisms.

We implemented this approach with a software frameworkcalled Synesketch.[1] Since emotion recognition software is rarely available online, we consider as our important additional contribution that Synesketch is an entirely free open-source library. Before Synesketch, there was only one such library with emotion recognition features—ConceptNet[2] [48]. Nevertheless, the latest and current version of ConceptNet does not include that feature.

Several Synesketch-based projects and applications have been developed, both by the authors and by third-party developers and designers, underpinning the practical and creative value of our contribution.

We have conducted an evaluation study of the recognition algorithm, which gave us promising results. In particular, the results show high classification accuracy and underline the importance of the emoticon lexicon. The study method and results are presented and discussed in detail in the Section 6 and Section 7.

- *U. Krcadinac, J. Jovanovic, and V. Devedzic are with the Department of Software Engineering, Faculty of Organizational Sciences, University of Belgrade, 11000, Jove Ilicaa 154, Belgrade, Serbia. E-mail: uros@krcadinac.com, {jeljov, devedzic}@gmail.com..*
- *P. Pasquier is with the School of Interactive Arts and Technology, Simon Fraser University, 250-13450 102nd Avenue, Surrey, BC V3T 0A3, Canada. E-mail: pasquier@sfu.ca.*

Manuscript received 10 Oct. 2012; revised 9 Aug. 2013; accepted 28 Aug. 2013; published online 5 Sept. 2013.
Recommended for acceptance by R. Mihalcea.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2012-10-0093.
Digital Object Identifier no. 10.1109/T-AFFC.2013.18.

[1] Synesketch, a textual emotion recognition and visualization software. Available at http://krcadinac.com/synesketch.

[2] http://conceptnet5.media.mit.edu/.

1949-3045/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

An additional feature of the Synesketch is a dynamic, emotion-related, generative art engine. We carried out an evaluation study of the emotion visualization system, but it is out of the scope of this paper. Interested readers are kindly referred to [4] to learn more about it.

2 Related Work

Problems of emotion recognition, representation, and evocation by computers have been widely investigated within the field of affective computing [5]. Approaches to textual emotion recognition and classification could be categorized into spotting emotional keywords, statistical language processing, and approaches based on large-scale general-level knowledge. Techniques also differ according to the domain and intended applications. However, due to intrinsic semantic ambiguity and imprecision of both human affects and texts that express them, none of the existing methods is perfect; each one has its pros and cons.

A research area closely related to emotion recognition (i.e., *affect analysis*), but distinct is *sentiment analysis*. Both of these areas focus on emotion recognition, but they use different models of representation and should not be equated. *Sentiment analysis* refers to the problem of positive/negative valence classification ([6], [7], [8], [32]), whereas *affect analysis* is about more fine-grained emotion recognition.

Since our focus is on classifying text in six emotion categories and determining actual emotional weights, our research primarily belongs to the latter field. However, since it also estimates the emotional valence, our approach *could* also be used for sentiment analysis, but it was not originally designed for that purpose. Hence, in what follows we focus primarily on affect analysis research.

2.1 Affect Lexicons and WordNet

Most textual sentiment and affect recognition research includes building and employing lexical resources with emotional keywords, i.e., words typically associated with certain emotion types.

Especially relevant to our work are those approaches that use WordNet, a lexical database for English language that also contains semantic connections between words [3]. So far, WordNet has been primarily used for sentiment analysis. For example, Esuli and Sebastiani [9] created a SentiWordNet lexicon based on WordNet synsets collected from synonymous terms. Three numerical weights, defining to what degree the terms are positive, negative, or neutral, were associated with each WordNet set of synonyms. Similarly, Andreevskaia and Bergler [10] presented a method for extracting sentiment-bearing adjectives from WordNet and assigning them positive or negative tags. Neviarouskaya et al. [11] described a WordNet-grounded method to automatically generate and score a sentiment lexicon, called SentiFul, and expand it through direct synonymy, antonymy, and hyponymy relations, derivation, and compounding with known lexical units.

In the field of emotion recognition, WordNet was used for creation of fine-grained emotion lexicons. For example, Strapparava and Valitutti [12] developed WordNet-Affect, a lexicon of affective concepts, based on a subset of WordNet set of synonyms. Affective labels for the concepts related to emotional state, moods, traits, situations evoking emotions or emotional responses were assigned to the WordNet-Affect entries. Strapparavaet al. [13] extended WordNet-Affect with a set of hierarchically organized emotional categories. Nevertheless, this organization is only partially compliant with the Ekman classification. For example, it includes labels such as "apprehension" (negative emotion), "anticipation" (positive), or"apathy" (neutral), which cannot fit well into Ekman's scheme. Ma et al. [14] searched WordNet for emotional words for all six emotional types defined by Ekman, and assigned to those words weights according to the proportion of synsets with emotional connotation those words belong to.

2.2 Keyword-Spotting Approach

A traditional and the most intuitive approach to emotion recognition is based on spotting emotional keywords. While some techniques take into account only words, others associate words with certain numerical weights. The Affective Reasoner by Elliott [15], being one of the simple approaches, searches for emotional keywords in text and uses a small lexicon of unambiguously affective words. Boucouvalas and Zhe [33] apply a more complex language parser in conjunction with a tagged dictionary of common words. Subasic and Huettner [17] associate words with numerical weights, grounding the method in fuzzy logic. Similarly, Ma et al. [14] use a lexicon with numerical weights in conjunction with sentence-level processing. Chuang et al. [49] also use a weighted lexicon with a simple rule-based system, not taking emoticons into account. Being the most popular approach, keyword spotting could also be found in the work of other researchers: Olveres et al. [18], Devillers et al. [19], Strapparava and Valitutti [12], Tao and Tan [20], Andreevskaia and Bergler [10], Wensel and Sood [21], and Francisco and Gervas [53].

In the realm of sentiment analysis, expecially interesting is a keyword-spotting method used by Thelwall et al. [47], who use emoticons and heuristics. However, their approach is limited to three emotional types (positive, negative, and neutral) and is heavily grounded on human coder subjective judgments.

Although keyword-based methods are praised for their intuitiveness, accessibility, and economy, they have been criticized for being based on shallow analysis capable of recognizing only surface features of the prose, and ignoring many semantic subtleties (Liu et al. [22], Seolet al. [23], Neviarouskaya et al. [24]). For example, these methods can fail to account for negation and rely on keyword sets which may be difficult to define. Moreover, they have problems with word sense disambiguation.

2.3 Statistical Approach

The most common alternative to keyword spotting is based on statistics and the use of machine learning algorithms trained with a large textual corpus. For example, Alm ([25], [50]) used supervised machine learning with the SNoW learning architecture in order to classify emotions in text. As a domain, the authors use children's fairytales. Furthermore, Aman and Szpakowicz [26] utilize a machine learning model with corpus-based features (unigrams). As a corpus, they use a collection of blog posts. Katz et al. [27] also employ a supervised system based on a unigram model. Similarly, Strapparava and Mihalcea [28] uset he Latent Semantic Analysis (LSA) technique and a Naive Bayes classifier trained on the corpus of blog posts annotated with emotions. Purver and Battersby [43] and Yuan and Purver [44] utilize emoticons as classification labels for the emotion classification task. Other researchers who used statistical language modeling techniques to analyze moods in text include Mishne [29], Leshed and Kaye [30], Mihalcea and Liu [31], and Calix et al. [49]. Keshtkar [52] opted for a hybrid approach that combines keyword-spotting with statistical approaches.

However, this alternative to keyword spotting has problems too: the lack of semantic precision (as opposed to keyword-based approaches), large corpora needed for solid performance, and, more often than not, neglect of negation and other syntactical constructs ([22], [24]).

Statistics-grounded techniques are also popular within the related field of sentiment analysis and classification ([32], [7], [51], [57], [58]). As previously said, sentiment analysis seeks to determine if a piece of text has a positive or a negative connotation. Especially interesting is the approach by Read [6], which utilizes several positive and negative emoticons in Usenet groups in order to train their software. Similarly, Carvalho et al. [8] use smiling emoticons and expressions such as "LOL" as one of the strategies for detecting irony in text. Go et al. [45] and Pak and Parouback [46] also use emoticons in order to determine the sentiment of the text.

2.4 Ruled-Based and Hybrid Approaches

There are other approaches, such as advanced rule-based linguistic approaches targeting textual affect recognition at the sentence level (e.g., [16]). Liu et al. [22] introduce an original approach based on a large-scale common sense knowledge base. The authors ground their affect models in Open Mind Common Sense, an open-source database of general common sense knowledge. This approach is implemented through ConceptNet, an open source toolkit [48]. Even though this approach is innovative and seems to offer a lot of potential, the presented evaluation results do not provide enough evidence of this technique's emotion classification accuracy. Rule-based systems are also employed in the field of sentiment analyses ([34], [35]).

Furthermore, there are some hybrid approaches to emotion recognition. For instance, Seolet al. 23] propose a system that consists of both a keyword recognizing engine and an emotion classifier. The classifier employs Knowledge-Based Artificial Neural Network (KBANN), which uses approximate domain knowledge and rules.

Finally, Neviarouskayaet al. [24] propose a rule-based linguistic approach for affect recognition from text, called the Affect Analysis Model (AAM). The authors employ a lexicon that consists not only of words, but also of emoticons and informal language. However, unlike the majority of approaches that rely on Ekman's six types of emotions, their model introduces nine emotional categories.

2.5 Summary

None of the techniques proved to be fully effective. The reported accuracy levels of affect analysis techniques (acquired using different methods and datasets) include: 45-65 percent [23], 68.2-79.9 percent [24], 54-79 percent [25], 71.3-73.9 percent [26], 56-81.5 percent [42], 62.8-85.9 percent [44], 67-70.5 percent [49]. Sentiment analysis techniques, such as [8], [45], and [47], report accuracies between 60 percent and 90 percent.

We decided to tackle the problem in the context of our domain of interest: massive online everyday communication structured in small textual fragments, such as online comments, chat conversations, tweets, etc. We propose a technique we believe is appropriate for the domain: a keyword spotting method enhanced with WordNet-based affinity lexicon, hand crafted heuristic rules, and taking into account emoticons, abbreviations, and a bulk of informal language in general.

Differences between our and related approaches are presented in the Appendix table (in the supplemental material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2013.18, and on our web site[3]). We analyzed different approaches in terms of

1. classification type,
2. classification method,
3. method features (inclusion of emoticons/abbreviations, based on rules, based on WordNet),
4. availability of test data, and
5. applications (free software, available online, open source, has real-world apps based on it, has visualization).

In terms of (5), Synesketch and the affective component of ConceptNet [48] are, to our knowledge, the only entirely free open source libraries for affect analysis. However, the current version of ConceptNet[4] does not include the affective component. The works of Purver and Battersby [43] and Yuan and Purver [44] are available online with limited features. Their API is not open source and only the basic version of the software is free. In the field of sentiment analysis, there are more available APIs and libraries. For example, [45] and [47] are available online. The system in [45] is not free; [47] has a free basic version.

3 Textual Emotion Recognition

Our hybrid keyword spotting technique is based on a lexicon and several heuristic rules. The lexicon consists of two parts: 1) a word lexicon, and 2) an emoticon lexicon. The word lexicon is semi-automatically generated using the WordNet lexical database for English language [3].

The emoticon lexicon, which also includes common abbreviationsand informal language common in Netspeak, is constructed manually. Each lexicon entry (word or emoticon) is associated with six emotional weights that correspond to six basic emotional categories defined by Ekman [2]: happiness, sadness, anger, fear, disgust, and surprise. The value of each weight is between 0 and 1. We opted for Ekman's model since it is the most common in the

[3] http://krcadinac.com/synesketch/researcj/tac.syn.appendix.pdf.

[4] http://concept5.media.mit.edu/.

field of emotion classification. Some researchers, such as Alm [50] and Calix et al. [49], refer to these emotional types as the "Big Six." As shown in the Appendix table (available online in the supplemental material and on our web site[5]), more than half of the presented emotion classification techniques use the Ekman's model.

3.1 Word Lexicon

The technique we use to generate the word lexicon is based on a simple idea proposed by Ma et al. [14]: The emotional weight of a word taken from WordNet can be calculated as a proportion of emotional senses among all senses of the word.

First, we start with a small initial collection of unambiguously emotional words and use it as a starting point for collecting the lexical "relatives" of these words from WordNet. The assumption of our technique is that words semantically close to this initial set of emotional words themselves carry a stronger emotional connotation than other words. So, in order to create the initial collection, we conducted a 20-person study. People were asked to list for each emotion type at least five words that they unambiguously associate the most with the given type. Words that were mentioned three or more times were considered good indicators of the corresponding emotion type and were added to the collection of words for that type. Such words were, for example, "happy" or "beautiful" for happiness, "lonely" for sadness, "terror" for fear, "rotten" for disgust, "suddenly" for surprise, etc.[6]

Then, we used WordNet 1.6 to search for synsets of the words from the initial collection. A synset is a set of semantically equivalent words within WordNet. Since most words carry more than one meaning, they belong to several synsets. Our lexicon is created through the analysis of semantic relationships of words and synsets, as described below.

The lexicon generation algorithm consists of the following steps:

1. six (empty) sets of emotional synsets S_k, $k \in E$, and the (empty) word set W are created. E is the set of six emotional types (happiness, sadness, anger, fear, disgust, surprise); $E = \{h, sd, a, f, d, su\}$.
2. WordNet is searched for synsets of words from the initial sets of emotional keywords V_k, $k \in E$. These initial synsets are added to S_k, $k \in E$, sets of emotional synsets for a given emotional type k.
3. This step is repeated d times. In each iteration l ($l = 1, 2, \ldots, d$), WordNet is searched for synsets semantically akin to the synsets from the S_k, via WordNet's pointer type *SimilarTo*. The extended synsets are added to S_k, $k \in E$. However, since these synsets are obtained indirectly, they are attached a penalty coefficient p, which is computed in the following manner:

$$p_{kj} = 0.1 * l;\ k \in \{h, s, a, f, d, su\}, j = 1, 2, \ldots, q_{ki}. \tag{1}$$

q_{ki} is the number of emotional synsets for the given word i and the given emotional type k (the emotional synset of type k is the one contained in the set S_k). The penalty grows in each iteration, which corresponds to the intuition that synsets semantically closer to the initial set of emotional keywords carry a stronger emotional meaning. In practice, the value of d is 3. The value of 0.1 (in (1)) and $d = 3$ resulted from a series of lab experiments that we have conducted on the test corpus of five sentences for each emotional type (not the corpus we used for evaluation). We varied this number and discussed the results with students and fellow researchers, and finally agreed on these values.

TABLE 1
A Small Portion of the Word Lexicon, with Emotional Weights Given for Several Words; Emotional Types Include: Happiness (H), Sadness (Sd), Anger (A), Fear (F), Disgust (D), and Surprise (Su)

Word	H	Sd	A	F	D	Su
joyful	1.0	0.0	0.0	0.0	0.0	0.0
severe	0.0	0.133	0.133	0.5	0.133	0.0
fierce	0.0	0.0	0.75	0.2	0.2	0.0
popeyed	0.0	0.0	0.0	0.0	0.0	0.45
repellent	0.0	0.0	0.0	0.0	0.25	0.0
somber	0.0	0.4	0.0	0.4	0.0	0.0

4. When all synsets are acquired, words from the synset sets S_k, $k \in E$, are added to the final set of words, W. The total number of words in W is m.
5. The emotional weights w_{ki}, $k \in E$, $i = 1, 2, \ldots, m$, are calculated for each word from W. For each word, the algorithm collects all the synsets in WordNet the word belongs to. For a given word i, the number of all synsets from WordNet is n_i. Some of these synsets also belong to the synset sets S_k—those are considered the emotional ones. Other synsets, though being part of WordNet, do not belong to the sets of emotional synsets. The emotional weight for each word and for each emotional type is calculated as a quotient between the number of emotional synsets (of a given emotional type) and the number of all synsets the word belongs to, diminished by using the average penalty of all its emotional synsets. This can be formally expressed in the following manner:

$$w_{ki} = \frac{q_{ki}}{n_i}\left(1 - \frac{\sum_{j=1}^{q_{ki}} p_{kj}}{q_{ki}}\right) = \frac{1}{n_i}\left(q_{ki} - \sum_{j=1}^{q_{ki}} p_{kj}\right) \tag{2}$$
$$i = 1, 2 \ldots, m;\ k \in \{h, s, a, f, d, sp\}.$$

The word lexicon formed this way consists of 3,725 words. A small part of the lexicon is shown in Table 1.

3.2 Emoticon Lexicon

An emoticon is a typographical symbol (or a combination of symbols) that represents a facial expression, such as :), ;), and the like. By using emoticons, writers tag their sentences with a certain emotion or mood, indicating in a more explicit way how the sentence should be interpreted. The idea of an emoticon-like symbol is actually older than the

[5] http://krcadinac.com/synesketch/research/tac.syn.appendix.pdf.

[6] The full list of seed words obtained by the 20-person study: http://krcadinac.com/synesketch/research/syn.seed.word.pdf.

Internet itself; for example, Ludwig Wittgenstein debated the power of a face-like symbol drawn by only four hand-drawn strokes to express a wide range of emotions [36]. Emoticons arguably express human feelings more directly than words [37].

Considering the widespread use of these symbols, we strongly argue that any textual sensing algorithm that focuses on online communication (such as chat) should consider emoticons. Unfortunately, emoticons are not part of WordNet or any other lexical database that we know of. Therefore, it was not possible to create an emoticon lexicon automatically—we had to do it manually. So, we first collected the most frequent text-based emoticons from the list of emoticons used by the most popular chat systems: GTalk,[7] Skype,[8] MSN Messenger,[9] and Yahoo! Messenger,[10] appended by the Wikipedia list of Western emoticons[11] as well as the list created by Gajadhar and Green [37]. Our list also includes common abbreviations, such as "LOL" or "OMG." In addition, the emoticon lexicon is appended with the common vulgarisms or informal exclamations ("damn" or "yuck," for instance), which though not emoticons, do not exist in lexical databases, yet carry an undeniable affective connotation.

Although the emoticons' emotional semantics is most often obvious and self-explanatory, we consulted Wikipedia's definitions of emoticons' emotional meanings. We consider this source relevant in this particular context since it is the result of a social consensus among many active Web users. However, Wikipedia descriptions do not comply with Ekman's categories for all emoticons from our collection, so we have conducted a study in order to define emotional vectors (with an emotional weight for each of the six emotional types) for each emoticon, as that information was not available. In order to ground the opinion in the common viewpoint, we contacted 174 participants on popular social networks using the Snowball sampling technique. All participants use social networks and chat regularly. Their ages vary between 16 and 40 years. Participants were asked to assign emotional weights to emoticons taken from the collection according to their perception of the emotions expressed by those emoticons. The choice of values for weights was: 1.0 (direct emotional meaning, the way happiness is associated with ":-)"), 0.5 (partial emotional meaning, the way fear, sadness, or disgust may (or may not) be associated with ":/"), or 0 (no emotional meaning). After the study, we assigned to the each emoticon a majority weight given by participants. A small portion of the emoticon lexicon is presented in Table 2. The entire emoticon lexicon consists of 128 symbols and abbreviations. Both word and emoticon lexicons are available online.[12]

3.3 Emotion Recognition and Heuristic Rules

In a nutshell, our recognition algorithm gets one sentence as its input, parses the text into words, compares these words with the ones from the lexicons, and then employs several heuristic rules. Parsing is done using the Java *BreakIterator* class.[13] In the process of text parsing, emoticons are used in the following way: If an emoticon is followed by a word with the first letter in uppercase, the word is considered the beginning of the next sentence.

TABLE 2
A Small Portion of the Emoticon Lexicon

Emoticon	H	Sd	A	F	D	Su
:-)	1.0	0.0	0.0	0.0	0.0	0.0
>:-(	0.0	0.5	1.0	0.0	0.5	0.0
lol	1.0	0.0	0.0	0.0	0.0	0.0
yuck	0.0	0.0	0.0	0.0	1.0	0.0

Heuristic rules, grounded in common sense, are intended to overcome some of the problems associated with keyword spotting methods, such as negation detection, the effect of punctuation marks, etc. Finally, the algorithm calculates the overall emotional state for the input sentence. The overall emotional state consists of an overall vector with six emotional weights and an emotional valence. The emotional valence can take values of -1, 0, or -1, showing if the emotion is negative, neutral, or positive, respectively.

The sentence-level rules, which apply to a whole sentence, are the following:

a. If there is a negation[14] in a part of a sentence (divided from the rest of the sentence by a comma, semicolon, colon, dash, or hyphen) where an emotional word is spotted, the emotional valence of the whole sentence is flipped. It means that the values switch between the happiness weight and the weights of the negative emotions (sadness, anger, fear, or disgust). For example, the algorithm would flip the valence of the sentence "I am not happy, that was hard," but would not flip the valence of the sentence "I am happy, that was not easy." Moreover, if the valence changes from positive to negative, the algorithm would assign the happiness weight to all four negative emotions (sadness, anger, fear, and disgust). The approach to handling negation in terms of its use of clausal punctuation could be found, for example, in [57] and [58]. However, the current version of our algorithm has certain limitations (for example, flipping all four negative emotions and handling double negation). We are currently working on a new version of the algorithm in order to fix these limitations (see Section 7 for details).

b. The more exclamation marks ("!") a sentence has, the more intense its emotions become. For each new mark, emotional weights get intensified by 20 percent.

c. If a sentence possesses a combination of characters such as "?!" or "!?", there is an emotion of surprise in it (the surprise weight is set to 1.0).

[7] http://tinyurl.com/gogle-list-of-emoticons.

[8] http://www.skypeemoticonslist.com.

[9] http://messenger.msn.com/Resource/Emoticons.aspx.

[10] http://messenger.yahoo.com/features/emoticons.

[11] http://en.wikipedia.org/wiki/List_of_emoticons.

[12] Complete open source of the Synesketch project: http://www.synesketch.krcadinac.com/download/synesketch.zip. Complete Java documentation of the Synesketch project: http://synesketch.krcadinac.com/javadoc/.

[13] http://docs.oracle.com/javase/6/docs/api/java/text/Breakiterator.html.

[14] Words used to detect negation include: *no, not, don't, haven't, weren't wasn't, didn't, hadn't, shouldn't, wouldn't*, and *won't*, together with the version of the same words without the apostrophe.

Word-level rules, which apply to single words, are the following:

d. The more characteristic signs a spotted emoticon has, the more intense the emotions of that sentence become. For example, this emoticon ":))))" is clearly more "happy" than this one ":)". For each new mark, related emotional weight (in this case happiness weight) gets intensified by 20 percent.

e. If a spotted emotional keyword is uppercase, the emotion associated with the word gets intensified by 50 percent.

f. If a spotted emotional keyword is preceded by an intensifying word (such as "extremely," "very," "exceedingly," etc.), the emotion associated with the word gets intensified by 50 percent. The values of 20 percent and 50 percet resulted from a series of lab experiments mentioned in the Section 3.1 (test corpus of five sentences for each emotional type).

Our algorithm consists of the following steps:

1.1 The input sentence is processed by applying sentence-level rules: a, b, and c.
2.2 The input sentence is parsed into words.
3.3 Each word is compared to keywords from both lexicons.
4.4 If a keyword is spotted, word-level rules—d, e, and f —are applied to it.
5.5 Emotional weights of a keyword are updated based on the applied (word-level) rules.
6.6 The keyword is added into an emotion words set. This is done for all spotted emotion-related keywords.
7.7 The overall emotional state of the sentence, the overall vector that corresponds to the entire sentence, is calculated using the emotion words set with updated weights.

Emotional weights of the overall vector are based on the max value of all keywords of the same emotion type from the emotion words set. Emotional valence depends on whether or not the sum of overall happiness weight outweighs the overall weight of the dominant negative emotion (sadness, anger, fear, or disgust weights).

Let the value ws_k denote the overall emotional weight and the value v the emotional valence for a given sentence and for a given emotional type k, $k \in \{h, sd, a, f, d, su\}$. The given sentence contains m emotional words. Let the value w_{ki} denote the emotional weight for a given word i, $i = 1, 2, \dots, m$, and a given emotional type k. Then, the overall emotional weights and the emotional valence of the sentence can be calculated in the following manner:

$$ws_k = max(w_{ki}); i = 1, 2, \dots, m;\; k \in h, sd, a, f, d, su\}$$

$$v = \begin{cases} -1, & w_{hi} - max_{u \in \{sd,a,f,d\}}(w_{ui} < 0 \\ 0, & w_{hi} - max_{u \in \{sd,a,f,d\}}(w_{ui}) = 0\,; i = 1, 2, \dots, m. \\ 1, & w_{hi} - max_{u \in \{sd,a,f,d\}}(w_{ui} > 0 \end{cases} \tag{3}$$

As an example, we will analyze the sentence "I won't be lovesick!" First, the algorithm would recognize an emotional keyword "lovesick", which has the following emotional vector: [0, 0.9, 0, 0, 0, 0], carrying only the sadness weight of 0.9. However, there is a negation ("won't" as "will not") in the part of the sentence where the emotional keyword is, so the valence becomes flipped (sentence-level rule, a): The happiness weight takes a value of the dominant negative weight, which is in this case sadness. Moreover, because the sentence ends with one exclamation sign, the value of happiness weight becomes intensified by 20 percent (sentence-level rule, c). The integral emotional vector of the whole sentence is: [1, 0, 0, 0, 0, 0]. The emotional valence is 1.

4 Applications

Our affect sensing approach is implemented through a textual emotion recognition engine called Synesketch [38]. In addition to the emotion recognition, Synesketch also provides software modules for emotion visualization in the form of abstract generative animation art. Synesketch is written in Java. Its first version was published online as a free open-source project (under the GNU General Public License) in November 2008. The library was improved over time. The version we describe in this paper is the final version downloadable from the website.

The goal of the visualization is to foster and expand the means and forms of human on-line communication and expression, by not only communicating emotions, but also evoking emotions in the users.Synesketch may be used in a variety of contexts, from market research based on fine-grained emotion analysis, through e-learning, to creative applications used for making user experience more enjoyable and fun. In fact, Synesketch has already been integrated into a couple of real-world apps, as shown in Section 4.2.

Synesketch won an award from the Belgrade Chamber of Commerce[15] and the International Digital Media and Arts Association (Simon Fraser University, Student Showcase Award).[16]

4.1 Synesketch Visualization

We developed two default visualization systems. Both of them transfer a sequence of sentences (written during a chat session, for example) into a generative animation. Each sentence triggers one type of animation, which is active until suppressed by the animation created for the next sentence. Generative animation art represents recognized emotions using a variety of color palettes, shapes, sizes, frame rates, and animation properties. Images of one of the visualizations are shown in Fig. 1. Animation examples of both visualizations are available online.[17]

Synesketch color palettes are based upon color combinations, specific for each emotional type, proposed by Da Pos and Green-Armytage [39]. These authors state that despite the subjectivity of the color-emotion mapping, most people, for example, tend to associate happiness with warm vivid colors and sadness with desaturated ones.

In order to assess the performance of our generative visualization in terms of user experience, we organized a special evaluation study. That study exceeds the scope of this paper, but is explained in detail in [4]. The evaluation results

[15] www.kombeg.org.rs/aktivnosti/komora/Komora.aspx?veza=259.

[16] http://idmaa.org/conferences/past-conferences/award-winners/.

[17] www.youtube.com/watch?v=u5kznE6kYmc.

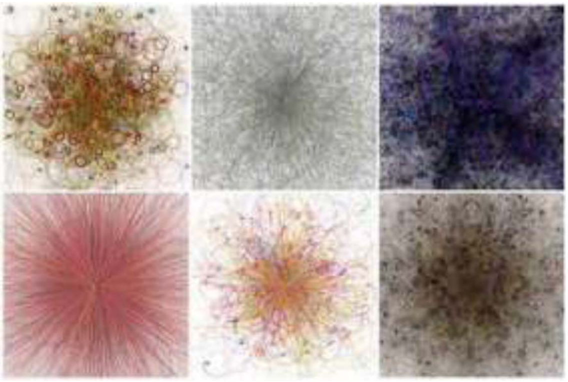

Fig. 1. Examples for one of the default Synesketch visuals representing six emotion types: first row: happiness, fear sadness; second row: anger, surprise, disgust. Generative art is partially based upon the Bubble Chamber, a work by artist Jered Tarbel (www.levitated.net/p5/chamber/).

justify our choice of visuals. For instance, the results show that Synesketch visualizations are highly effective in communicating emotions and, compared to other common emotion visualization techniques (specifically, animated chat emoticons and avatars), they are (statistically) significantly betterin evoking emotions. However, these visualizations are only the default ones; Synesketch supports and promotes building visualizations by third-party designers.

4.2 Synesketch-Based Projects

The Synesketch software was built following a good practice of object-oriented software design, with a heavy use of design patterns [41]. The visualization engine and the recognition engine are completely independent. This architecture allowed both us and third-party Synesketch developers and designers to create different kinds of extensions and standalone software based on our library.

We built the emotional visual chat as a software extension of the Skype software, one of the most widely used applications for online communication; the software is called SyneSkype.[18] SyneSkype adds one new feature to the standard Skype chat: a window for visualizations of emotions detected in the exchanged chat messages.

Our second application grounded on Synesketch was a module for Moodle, a widely used Learning Content Management system; this module allows for following the affect of students' interaction [54]. We used the emotion recognition feature in conjunction with appropriate visualizations of students' affective interaction to provide teachers with timely and easy-to-use feedback that can be leveraged to coordinate the learning process.

Third-party systems include, for example, software for analysis and visualization of tweets, such as EmoTweet[19] and Twitter Emotion Graphs (TEG).[20] There is also Synesketch-based lyrics-visualizing Karaoke software.[21] Synesketch was featured in Infosthetics.com[22] and VisualComplexity.com.[23] The Creative Review magazine (2008) applied Synesketch visualizations to the poetry of John Hegley and Sylvia Plath. Synesketch visualizations were also applied to news websites,[24] and to comments from the source code of various software projects in order to show how programmers feel about their software.[25]

5 Evaluation

In order to assess the performance of our emotion recognition algorithm, we organized an evaluation study that was driven by the following research questions

RQ1—Is the performance of the proposed recognition algorithm comparable to human performance in the evaluation of emotional value of sentences? In other words, we wanted to explore the similarity between emotional ratings of sentences generated using the proposed algorithm and human ratings of the same set of sentences.

RQ2—Do both the word lexicon and the emoticon lexicon significantly impact the performance of the proposed recognition algorithm?

We stress that determining a border between emotional and neutral sentences was not in the scope of these research questions. The distinction between neutrality and emotionality tends to be especially fuzzy and prone to human disagreement (as discussed, e.g., in Alm's study [50]). The intention of this study was to assess, if a sentence is emotional, whether emotional types and weights were properly recognized.

In order to compare our approach with concurrent ones and determine the accuracy of our method in terms of detection of neutrality in text, we organized an additional experiment explained in Section 6.

5.1 Design

Using a corpus of 149 English sentences, we have computed the following metrics: a human gold standard and four computer-generated metrics. To obtain the metrics, sentences were rated in terms of emotional type and emotional weights by 1) human participants (a human gold standard), 2) Synesketch algorithm with the emoticon lexicon alone, 3) Synesketch algorithm with the word lexicon alone, 4) Synesketch algorithm with both lexicons, and 5) the most widely accepted algorithm for generating random numbers within the range 0-1, based on linear congruent generators [55]. Since human metrics are inherently difficult to randomize, this algorithm was introduced as a basic reference for comparison purposes only, and not to closely reflect human metrics.

The gold standard was based on average ratings given by humans. In order to answer RQ1 and RQ2, we did three kinds of comparisons. First, for each sentence from the corpus, we compared the human gold standard on one side with each computer output on the other side. In particular, we calculated Pearson Product Moment Correlation Coefficients (PMCCs) and cosine similarity between the gold standard and each of the computer-generated metrics.

Second, we compared the Emotional Valence (EV) accuracy and the Dominant Emotional Type (DET) accuracy. With the human gold standard as a measure for

[18] http://synesketch.krcadinac.com/blog/?page_id=21.
[19] http://dhairyadand.com/works/emotweet/.
[20] http://davidguttman.com/twitter_emotion_graphs.
[21] http://goo.gl/PtAZ75.
[22] http://go.gl/s5QCS8.
[23] http://www.visualcomplexity.com/vc/roject_cfm?id=695.
[24] http://goo.gl/0cVOk8.
[25] http://www.natpryce.com/articles/000748.html.

correctness, the EV accuracy for a given metric is a percent of sentences with the correctly recognized valence. Similarly, the DET accuracy for a given metric is a percent of sentences with the correctly recognized emotional type.

Third, for comparison of the emotional weights we used absolute values of the distance between the human ratings and the computer-generated ones. The smaller the distance between the outputs, the better the algorithm performed in our tests.

5.2 Participants

The above mentioned sentence corpus was presented to the study participants who were asked to rate the emotional type and intensity of each of the sentences. The participants were recruited at the University of Belgrade from an undergraduate course in programming and computer science. All of the participants were fluent in English and used web-based communication tools regularly, as those were the requirements for participation in the study. Overall, 214 students completed all the study tasks.

5.3 Materials

We tested the proposed approach with a corpus of 149 English sentences. The first half of these sentences was randomly taken from Group Hug,[26] a Website that publishes anonymous confessions in the form of short textual posts. The large majority of texts published on this Website contain an obvious emotional connotation. We especially opted for Group Hug because sentences were anonymously written, thus being more honest and expressive in term of affects, especially negative ones.

The second half of the corpus was gathered through a small survey. Fifteen people (six female and nine male) were asked to write down several sentences for each emotional type—sentences they unambiguously associate with a certain emotional type. They were told to feel free to use any form of informal language, just as they would do it online. All of them are fluent in English and use computers and Web regularly. Their ages vary between 24 and 55. One should note that these participants only took part in the creation of the study corpus and did not participate in the (emotional) rating of sentences from the corpus. The complete list of 149 sentences is available online.[27]

5.4 Data Collection and Analysis

Each participant was given 20 sentences randomly taken from the corpus. They were asked to rate each sentence on six scales, one for each emotional type. Each scale contained values from 1 to 5, 1 referring to the absence of emotion, and 5 referring to the full emotion. We eventually ended up with each sentence being rated by, on average, 29 participants (min 15, max 58). Results are also available online.[28]

Once we collected the participants' answers, we mapped the results to [0, 1] to match the Synesketch algorithm outputs. Mapping discrete values to continuous emotional weights was done using the medium baseline: 1 was mapped to 0.0, 2 to 0.25, 3 to 0.5, 4 to 0.75, and 5 to 1.0. In order to acquire our gold standard, we calculated an average emotional vector (1 emotional weight for each emotional type) for each of the 149 human sentence ratings.

TABLE 3
Average Means (avg) and Average Standard Deviations (SD) among Human Raters for All 149 Sentences; Also, Number of Sentences for Each Emotional Type (NS) Where a Particular Emotional Type Is the Dominant One; Emotional Types Include: Happiness (H), Sadness (Sd, Anger (A), Fear (F), Disgust (D), and Surprise (su)

Mtr	H	Sd	A	F	D	Su
Avg	0.278	0.227	0.164	0.108	0.110	0.208
SD	0.108	0.156	0.141	0.131	0.108	0.16
NS	45	40	18	12	13	21

Table 3 presents the average means (first row) and average standard deviations (second row) among human raters for all sentences and for each emotional type. These values show the level of human's agreement on emotional bearings of the sentence corpus. Also, it presents, for each emotional type, the number of sentences dominated by that particular type (third row). For example, there are 45 "happy" sentences, 40 "sad" ones, and only 12 sentences where fear is the dominant emotional type.

The criterion for choosing a dominant emotional type was the maximum value of the vector, no matter how small. That is why there are no neutral sentences in the table.

After we defined the gold standard, we computed PMCCs and cosine similarities between the gold standard and each of the metrics (three Synesketch-based metrics and one random) for every sentence from the corpus. PMCC and cosine similarity were chosen because they are standard measures of correlation/similarity for interval data such as ours [56]. Also, we computed an absolute value of the difference between the human gold standard and each metric for every sentence from the corpus. The result was four Tables with six columns (one for each emotional type) and 149 rows (one for each sentence), where each cell contained the absolute distance between a given metric and the human gold standard for a given sentence and a given emotional type. Finally, we calculated the EV accuracy and the DET accuracy for all four metrics.

5.5 Results

To address our first research question (RQ1), we compared the human gold standard with each computer output, using three criteria: 1) PMCCs, 2) EV and DET classification accuracy, and 3) average differences between human and computer-based metrics.

5.5.1 *RQ1: Correlation between Human and Computer-Based Metrics*

Table 4 gives the PMCCs between the human gold standard and each of the four metrics. Table 5 gives, for each metric, the average values of the PMCCs across all emotional types. Higher PMCC indicate a higher similarity with the gold standard.

The results from Table 4 and Table 5 suggest that the Synesketch recognition algorithm with both lexicons—the

[26] http://en.wikipedia.org/wiki/Group_hug.
[27] http://goo.gl/D2pu8D.
[28] http://goo.gl/0fNr4r.

TABLE 4
Correlation (Pearson Product-Moment Correlation Coefficients) between Human Gold-Standard and Computer-Generated Metrics for All Emotional Types; Statistically Significant Correlations Are Given in Bold and Marked with * (p < 0.05, 2-Tailed) and ** (p < 1.001, 2-Tailed)

Mtr	H	Sd	A	F	D	Su
Rnd.	0.169*	-0.052	0.059	-0.164*	-0.067	-0.016
S.E.	0.519**	0.508**	0.565**	0.00	0.387**	**0.888****
S.W.	0.756**	0.565**	0.488**	**0.577****	0.696**	0.756**
Syn.	**0.889****	**0.780****	**0.727****	**0.577****	**0.756****	0.865**

TABLE 5
Average Pearson Product-Moments Correlation Coefficients (PMSS) for All Metrics across All Emotional Types

Metric Type	PMCC
Random	-0.0119
Synesketch Emoticons	0.478
Synesketch Words	0.64
Synesketch	0.766

TABLE 6
Cosine Similarity between Human Gold Standard and Computer-Generated Metrics for All Emotional Types

Mtr	H	Sd	A	F	D	Su
Rnd.	0.573	0.481	0.472	0.341	0.336	0.469
S.E.	0.574	0.555	0.578	0.00	0.406	**0.889**
S.W.	0.834	0.69	0.58	**0.664**	0.745	0.797
Syn.	**0.926**	**0.851**	**0.779**	**0.664**	**0.796**	0.888

TABLE 7
Average Cosine Similarity for All Metrics

Metric Type	Avg. cosine sim.
Random	0.445
Synesketch Emoticons	0.5
Synesketch Words	0.718
Synesketch	0.817

Synesketch metric (Syn)—outperforms other metrics on average and for all emotional types except fear and surprise.

For the emotion of fear, it seems that the emoticons play no role at all (Synesketch Emoticons (S.E.) metric has a PCMM equal to zero). The emotion estimation solely comes from the recognition of words. For the emotion of surprise, it seems that the emoticon lexicon alone works better than the combination of both lexicons (S.E. metric equals 0.888 while Syn. metric equals 0.865).

Table 6 gives cosine similarities between the human gold standard and each of the four metrics. Table 7 gives, for each metric, the average values of the cosine similarity across all emotional types. Higher cosine similarity indicates a higher similarity with the gold standard. Results from Table 6 and Table 7 confirm previous conclusions, particularly about the emotions of fear and surprise.

5.5.2 RQ1: Classification Accuracy

Table 8 provides the EV accuracy and the DET accuracy for all four metrics. If accuracy is higher, we acknowledge a better performance for a given metric. For additional comparison we employ a majority-class baseline (based on a negative emotion for EV and happiness for DET).

TABLE 8
The Emotional Valence (EV) Accuracy (acc.) and the Dominant Emotional Type (DET) Accuracy for Each Computer-Generated Metric, Random Baseline, and a Majority-Class Baseline

Metric Type	EV acc.	DET acc.
Majority-class	0.557	0.302
Random	0.543	0.154
Syn. Emoticons	0.255	0.322
Syn. Words	0.691	0.677
Synesketch	0.792	0.798

TABLE 9
Precison, Recall, and F-Measure for Each Emotion Type

	Precision (%)	Recall (%)	F-Measure
Happiness	0.851	0.889	0.869
Sadness	0.816	0.775	0.795
Anger	0.765	0.722	0.743
Fear	0.533	0.667	0.592
Disgust	0.769	0.769	0.769
Surprise	0.895	0.809	0.85

These results show notable differences among the three Synesketch metrics. Using the human gold standard as a measure for correctness, the results show that the Synesketch Emoticons metric recognized the correct valence in 38 and the correct dominant emotional type in 48 out of 149 sentences (25.5 percnet and 32.2 percent, respectively). This is drastically outperformed by the Synesketch Words metric: 103 sentences with the correct valence (69.1 percent) and 101 with the correct emotional type (67.7 percent). The Synesketch algorithm with both lexicons performs the best by recognizing the correct valence in 118 sentences (79.2 percent) and the correct dominant emotional type in 119 sentences (79.8 percent). This is an improvement of 23.5 percent and 49.6 percent over the majority-class baseline, for EV and DET, respectively.

However, although the random metric has the worst performance in recognizing the dominant emotional type among all the metrics (only 15.4 percent), it actually outperforms the Synesketch Emoticon metric in terms of correct emotional valence (0.543 compared to 0.255). This can be explained by the fact that the large majority of the sentences are not emotionally neutral, thus having the value of emotional valence either -1 or 1, which corresponds well to the random metric values.

We also calculated precision, recall, and harmonic F-measure for each emotional type separately (Table 9). It shows that the Synesketch algorithm shows the best results for emotions of happiness and surprise (F-measure being 0.869 and 0.85, respectively) and the worst result for the emotion of fear (F-measure being 0.592).

TABLE 10
Descriptive Statistics—Averages (a) and Standard Devatiations (sd)—of Absolute Differences between the Human Gold Standard and the Results of Each Considered Metric; the Statistics Are Given for Each Emotional Type and Are Based on All the Sentences from the Study Corpus; H, Sd, A, F, D, and Su Refer to the Six Emotional Types: Happiness, Sadness, Anger, Fear, Disgust, and Surprise, Respectively

Metric		**H**	**Sd**	**A**	**F**	**D**	**Su**
Rnd.	*a*	0.428	0.465	0.437	0.477	0.464	0.444
	sd	*0.284*	*0.288*	*0.279*	*0.280*	*0.284*	*0.293*
S.E.	*a*	0.199	0.182	0.129	0.108	0.1	0.095
	sd	*0.343*	*0.277*	*0.228*	*0.201*	*0.22*	*0.146*
S.W.	*a*	0.146	0.172	0.135	0.121	0.094	0.124
	sd	*0.235*	*0.248*	*0.229*	*0.187*	*0.175*	*0.203*
Syn.	*a*	0.103	0.136	0.107	0.121	0.0888	0.102
	sd	*0.167*	*0.19*	*0.177*	*0.187*	*0.162*	*0.156*

5.5.3 RQ1: Average Differences between Human and Computer-Based Metrics

Table 10 provides descriptive statistics (average values (*a*) and standard deviations (*sd*)) of absolute differences between the human gold standard and each of the four metrics. We report these values for each emotional type. The lower the average value is, the less distant the metric is from the gold standard. That is, if an average distance is smaller, we consider a better performance for a given metric.

Table 11 gives, for each metric, the average values of the average distances (reported in Table 10) across all emotional types. Table 11 suggests that, on average, the Synesketch recognition algorithm with both lexicons (the Synesketch metric) performs the best, that is, the average value of the absolute distance (0.11) is lower than for other metrics. This metric outperforms the Synesketch Emoticons metric (0.136) and the Synesketch Words metric (0.132), and by far the random metric (0.452).

If we look into the details, Table 10 shows that the Synesketch metric outperforms the rest for all emotional types too, except for the emotions of fear and surprise. Overall, all three Synesketch metrics show much smaller difference among themselves compared to the random metric, which has a significantly worse performance.

As a final response to RQ1, we can conclude that the performance of the proposed recognition algorithm (the Synesketch algorithm with both lexicons) has proven to be significantly better, i.e. closer to the gold standard, than the random metric. This is true for absolute distances between emotional weights, the EV accuracy, and the DET accuracy.

TABLE 11
Average Value of the Absolute Distances between the Human Gold Standard and the Results of Each Considered Metric

Metric Type	**Average Abs Distance**
Random	0.452
Synesketch Emoticons	0.136
Synesketch Words	0.132
Synesketch	0.11

TABLE 12
Paired t-Tests Scores for Comparison between Synesketch Metric and Synesketch Words Metric; Statistically Significant Effects Are Presented in Bold and with an Asterisk; Ms and Mw Refer to the Mean Values for Synesketch Metric and Synesketch Word Metric, Respetively; for the Emotion of Fear Test Could Not Be Performed Because the Values Are Zeros (No Difference between Ms and Mw Metric

	H	**Sd**	**A**	**F**	**D**	**Su**
Ms	0.103	0.136	0.107	0.121	0.089	0.102
Mw	0.146	0.172	0.135	0.121	0.094	0.124
t	**-2.813**	***-2.362***	**-2.155**	/	-0.661	**-2.017**
DF	**148**	**148**	**148**	/	148	**148**
p	***0.006****	***0.019****	***0.033****	/	*0.51*	***0.046****

5.5.4 Comparison between Synesketch Metrics

To respond to RQ2, we examined if the observed differences between Synesketch metrics are statistically significant. We used a paired (dependent) t-test to compare absolute differences between metrics. Paired t-test was chosen because we have a case of comparing the means of different metrics measured on the same set of subjects (i.e., sentences) [56].

Table 12 presents the results of a t-test that compares absolute differences of the results of the Synesketch metric and of the Synesketch Words metric. Similarly, Table 13 presents the results of a t-test that compares the Synesketch metric with the Synesketch Emoticons metric. Data are normally distributed and satisfy the assumptions of the t-test.

In the case of Synesketch metric vs. Synesketch Words metric, a significant difference ($p < 0.05$) was observed for the emotions of happiness, sadness, anger, and surprise. Since the absolute differences represent the distance from the gold standard, the results suggest that Synesketch metric performs better than Synesketch Words metric for these four emotional types.

In the case of Synesketch versus Synesketch Emoticons metric, a significant difference ($p < 0.05$) was observed only for the emotion of happiness. For all other emotional types no significant effect was found. Therefore, Synesketch metric is significantly better than Synesketch Emoticons metric only in the case of happiness.

So, in the context of R2, we can conclude that, in terms of emotional weights, both the word and the emoticon lexicon have a significant effect for the positive emotions.

TABLE 13
Paired t-Tests Scores for Comparison between Synesketch Metric and Synesketch Emoticons Metric; Statistically Significant Effects Are Presented in Bold and with an Asterisk; Ms and Me Refer to the Mean Values for Synesketch Metric and Synesketch Emoticons Metric, Respetively

	H	**Sd**	**A**	**F**	**D**	**Su**
Ms	0.103	0.136	0.107	0.121	0.089	0.102
Me	0.199	0.182	0.129	0.108	0.1	0.095
t	**-3.334**	-1.942	-1.566	0.730	-0.576	0.771
DF	**148**	148	148	148	148	148
p	***0.001****	*0.054*	*0.12*	*0.466*	*0.565*	*0.442*

5.6 Discussion of the Study Results

As expected, the proposed algorithm had a good performance for sentences such as *"He's too much fun!"*, *"I am not a happy one."*, *"Damn."*, or *"But all of a sudden I heard a cop yell."*. Sentences which proved to be problematic were the ones without visible emotional cues, such as *"Live your dreams!"*; sentences with complex and ambiguous emotional meaning, such as *"It's funny all the things I don't care about anymore"*; or sentences with expressions containing unusual sequences of characters, such as *"I am so full of it! they are #!$#%$!#~~!#!@~!"*.

Concerning the overall performance of the algorithm, we consider the EV accuracy and the DET accuracy to be the most valuable indicators of the algorithm's overall performance—for at least two reasons.

First, the "correctness" of emotional weights is arguably much more difficult to measure than the "correctness" of EV and DET. For example, one of the sentences from the corpus is: *"But it is all worrying."* Although most of the participants gave this sentence a dominant fear weight, the weight varied between 0.25 and 1. Accordingly, one could argue that the affective interpretations are more subjective in terms of weights than in terms of type and valence.

Second, the emotional type and valence are arguably more beneficial for the practical applications of the algorithm. Therefore, the finding that we consider especially important and useful is that our recognition algorithm was able to sense the correct emotional type and valence in ~ 80 percent of the cases. We thus believe that the algorithm has a practical value, which has also been demonstrated by the various Synesketch-based applications.

Another interesting finding is the difference between fear and surprise in terms of usefulness of emoticons. In the case of fear, it seems that emoticons play no role at all (see Table 4 and Table 6). That might be due to several reasons: 1) the small number of fear sentences (12 out of 149), 2) the small number of fear emoticons (13 out of 128), 3) the very nature of online communication, which has not developed the informal language for expressing fear as much as it developed it for, let's say, happiness. Contrastingly, in the case of surprise, emoticons seem to play a crucial role. There are more surprise emoticons and abbreviations (23 out of 128), but it also seems that the nature of the emotion of surprise is better expressed through emoticons, compared to other emotional categories. In any case, we believe this is an interesting topic for further research.

However, there are certain threats to validity that are worth mentioning. First, at least 50 percent of the evaluation corpus consisted of sentences with relatively unambiguous emotional type. This way, the evaluation possibly lost some of its generalization power, but, on the other hand, its results are clearer and easier to interpret. In the future, we plan to examine the algorithm in the wild, with a goal of improving it and making it more robust.

Second, not recognizing neutral as a separate type may have affected EV and DET accuracy because certain sentences with low emotional weights were not considered neutral. Participants involved in our study, both the ones who provided us with the 50 percent of the sentence corpus and the ones who rated the sentences, come from a similar background (students, teachers, or researchers at the University of Belgrade) and are mostly not native English speakers. However, these participants represent typical users of online communication channels and are quite familiar with the jargon of online communication. Moreover, the proposed approach is neither focused only on native speakers nor are any of its elements that it relies on (i.e., requires) perfectly written English sentences, so we do not consider this as a drawback. Therefore, in our opinion, the sentences and ratings they provided could be considered credible.

6 Experiment with Sentences from Fairy Tales

In addition to the main study, we compared the performance of the Synesketch algorithm with the affect recognition systems introduces by Alm [50] and Neviarouskaya et al. [24] on the corpus of sentences from fairy tales, defined by Alm. These systems reportedly outperformed ConceptNet affect recognition function (Liu et al. [22]). Alm states that ConceptNet proved to be worse than the baseline [50]. Apart from the comparison, this experiment was also aimed at determining the accuracy of our method in terms of detection of neutrality in text.

We followed the same evaluation method as Alm, utilizing three hierarchical levels of affect labels: level 1 (ALL) includes Ekman's emotional types plus neutral; level 2 (MID) uses three categories: positive, negative, and neutral; level 3 (TOP) includes two categories: emotional and neutral. The logic of the mapping was the same as the one we used when determining valence: Happiness was mapped to positive; sadness, anger, fear, and disgust were mapped to negative; surprise was neutral.

For the experiment, we used the subset of 1,207 sentences annotated as emotion bearing with high agreement (indicating that, for the particular sentence, emotional labels assigned by four human annotators were identical), and a subset of neutral sentences. Since we did not have the subsets of neutral sentences that Alm and Neviarouskaya et al. used in their experiments,we randomly extracted them from the whole corpus of sentences that was labeled by all human annotators as neutral.The number of affective labels at each hierarchical level determined the size of a sample of neutral sentences (taken from Alm 2008), as defined in the following equation:

$$\left[\frac{|HA|}{(|A_i|-1}\right]. \tag{4}$$

HA is the set of high-agreement affect sentences in the whole corpus; A_i is the set of affect labels at a specific level i in the emotional hierarchy (1, 2, and 3). This is because the balance between neutral sentences and other emotional ones was needed. For example, in the case of level 1, neutral is just one of the several emotional types. In the case of level 3, however, there are only two types, emotional and neutral, so half of the corpus had to consist of neutral sentences.

TABLE 14
Accuracy across Sentences from Fairy Tales in High Agreement; *i* Is the Level of the Emotional Hierarchy; # Is the Total Number of Sentences; *#N* Is the Number of Neutral Sentences; *n-bl* Is the Neural Baseline; *f-bl* Is the Most Frequent Emotion Baseline; *1*, *2*, *3*, and *4* Present the Individual Classification Methods; *1* Is the Alm's lextag Accuracy (Accuracy Is Given as a %), *2* Is the Alm's LOOHAsnowtag Accuracy (Span of Mean Accuracy), *3* Is the AAM Accuracy (Span of Accuracy across Three Algorithm Variations), and *4* Is Synesketch Accuracy

i	#	#N	*n-bl*	*f-bl*	1	2	3	4
1	1448	241	17	31	54-55	69-70	68.2-70.2	61.8
2	1810	603	33	40	60	69-73	73.3-75.5	65.6
3	2414	1207	50	50	69	79	77.6-79.9	75.6

Table 14 presents the results of the comparison. Accuracy is given for the three levels of Alm's emotional hierarchy and for the four classification methods:

1. Alm's lextag [50] (simple keyword spotting);
2. Alm's LOOHAsnowtag [50] (supervised machine learning);
3. the Affect Analysis Model (AAM) by Neviarouskaya et al. (rule-based approach) [24]; and
4. Synesketch.

We used two baselines, defined by Alm: *n-bl* is the ratio of neutrally labeled sentences; *f-blis* the ratio of the most frequent emotional label (for level 1, it was happiness; for level 2, it was the negative emotion). As Synesketch does not recognize neutral as a special type, we had to define a threshold for the dominant emotional type weight under which we would consider the sentence neutral. Using a sample of 100 sentences taken from Alm's dataset (not the sentences we used for the experiment), we determined a threshold of 0.2. If the maximum emotional weight in the vector was under 0.2, we considered the sentence neutral.

For all three levels, Synesketch's accuracy proved to be better than Alm's lextag, but somewhat worse than Alm's LOOHAshowtag and Neviarouskaya's AAM. However, because related studies did not present enough data, it is impossible to claim whether these differences are statistically significant. In addition, the results show that the accuracy is inversely proportional to the number of labels at the hierarchical levels (the highest accuracy is at the level 3 with only two possible categories).

Nevertheless, there are certain threats to validity of this experiment originating primarily in inherent differences of the compared tools. First, the method of choosing neutral sentences (taking a random sample from the entire corpus), which could not have been avoided, might have produced certain inconsistencies. Second, the nature of the neutral type itself might have affected the results. Synesketch does not treat neutral as a special type, so we had to determine a threshold. Third, the dataset combined the emotions of anger and disgust into one emotional type (label). Neviarouskaya et al. had a similar problem with certain inconsistencies in emotional models (Neviarouskaya et al. use nine basic emotional categories, including, for instance, the emotion of interest).

Most Synesketch errors come from the cases where additional context is required for correct interpretation of textual emotion. This emphasized the need to include word sense disambiguation in our future research.

Errors also occurred in situations when the correct emotion was in Synesketch's final vector, but based on its weight, it was not the dominant one. Further research is needed to determine how much the difference between the dominant and the second dominant emotion (difference sometimes very small) affects Synesketch's final accuracy.

The third kind of error is due to the vocabulary and style used in the language of fairy tales—which is, arguably, significantly different from the contemporary use of language (there are no emoticons, to name the most obvious difference). These errors might be reduced by expanding our lexicon, but it is uncertain whether this would actually make the algorithm sufficiently robust so that it can be taken out of the context of fairy tales and into the ″wilderness″ of the Web.

We wanted to do more comparisons with concurrent systems, but a large majority of datasets or programs were not publicly available. Even when available, other datasets are not comparable with our output: Some text corpus or programs were made for sentiment (not affect) recognition purposes. Moreover, the algorithms by other authors were not described in sufficient detail in order to be reimplemented. Hence, we had to rely on the results reported in related studies, and the only study that made a comparable dataset available was Alm's.

7 Conclusion and Future Work

We presented an approach for textual affect recognition positioned within the context of computer-mediated human interaction. Our recognition algorithm receives a text sentence as an input and classifies it according to the six emotional types defined by Ekman [2]. The output emotional vector can be used to determine the dominant emotional type and the emotional valence of the sentence. Our contribution is based on the way lexicons are created and how they are used in conjunction with heuristic rules. We use the WordNet-based word lexicon and a lexicon of emoticons, abbreviations, and vulgarisms.

Our affect recognition approach is illustrated through a software system, Synesketch. Besides the affect recognition engine, Synesketch includes an engine for generating emotion-related abstract animation in real time. Most of Synesketch-based applications, created by third-party developers and designers, are located within the context of computer-mediated human communication (such as emotional visual chat). Synesketch is, to the extent of our knowledge, the only free open source textual emotion recognition software published on the web.

The evaluation study presents promising results in terms of high classification accuracy, underlining the importance of the emoticon lexicon.

Future efforts will be centered on refining the algorithm and the lexicon. At the moment we are working on the improved version of negation detection. For example, instead of applying the positive weight to all four negative emotions, the new algorithm will try to determine the contextually relevant negative emotions and apply the positive weight only to them. In addition, the new algorithm will change the effect of modifiers if the negation is detected ("not very happy" does not imply an increase, but a decrease in weight). The new algorithm will also take into account a problem of double negation.

A challenging research question we plan to tackle is whether an algorithm with some "in-depth" natural language processing features would actually improve the recognition performance. For example, one of the weaknesses of our approach is that word sense disambiguation and part-of-speech assignment are not considered. We plan to add these features in a new release. New software features will also include users in the process of improving the lexicon in at least two ways. First, we will allow the users to evaluate emotion recognition while using Synesketch, for example, by giving a grade to the estimated emotional state of a sentence. The program would continuously follow this feedback and would adjust the lexicon accordingly. Second, the users will be allowed to add words, emoticons, abbreviations, etc., and annotate them with appropriate affects. The software would then search for recurrent patterns and add those to the lexicon. We also plan to expand our emoticon lexicon and give it a more formal semantics, through integrating the lexicon with the Smiley Ontology [40].

REFERENCES

[1] R. Peris, M.A. Gimeno, D. Pinazo, G. Ortet, V. Carrero, M. Sanchiz, and I. Ibanez, "Online Chat Rooms: Virtual Spaces of Interaction for Socially Oriented People," *Cyber Psychology and Behavior*, vol. 5, no. 1, pp. 43-51, 2002.

[2] P. Ekman, "Facial Expression and Emotion," *Am. Psychologist*, vol. 48, pp. 384-392, 1993.

[3] G.A. Miller, "WordNet: An On-Line Lexical Database," *Lexicography*, special issue, vol. 3, no. 4, pp. 235-312, 1990.

[4] U. Krcadinac, J. Jovanovic, V. Devedzic, and P. Pasquier, "Synesketch: Using Generative Animation to Communicate and Evoke Emotions Recognised in Text," unpublished. Available online: http://krcadinac.com/papers/synesketch_generative_animation/, 2013.

[5] R. Picard, *Affective Computing.* The MIT Press, 1997.

[6] J. Read, "Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classication," *Proc. ACL Student Research Workshop*, pp. 43-48, 2005.

[7] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1-2, pp. 1-135, 2008.

[8] P. Carvalho, L. Sarmento, M.J. Silva, and E.d. Oliveira, "Clues for Detecting Irony in User-Generated Contents: Oh...!! It's 'So Easy' ;-)," *Proc. First Int'l CIKM Workshop Topic-Sentiment Analysis for Mass Opinion*, pp. 53-56, 2009.

[9] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," *Proc. Fifth Int'l Conf. Language Resources and Evaluation*, pp. 417-422, 2006.

[10] A. Andreevskaia and S. Bergler, "Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses," *Proc. 11th Conf. European Chapter Assoc. for Computational Linguistics*, pp. 209-216, 2006.

[11] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Analysis of Affect Expressed through the Evolving Language of Online Communication," *Proc. Int'l Conf. Intelligent User Interfaces*, pp. 278-281, 2007.

[12] C. Strapparava and A. Valitutti, "WordNet-Affect: An Affective Extension of WordNet," *Proc. Int'l Conf. Language Resources and Evaluation*, pp. 1083-1086, 2004.

[13] C. Strapparava, A. Valitutti, and O. Stock, "The Affective Weight of Lexicon," *Proc. Fifth Int'l Conf. Language Resources and Evaluation*, pp. 423-426, 2006.

[14] C. Ma, H. Prendinger, and M. Ishizuka, "Emotion Estimation and Reasoning Based on Affective Textual Interaction," *Proc. First Int'l Conf. Affective Computing and Intelligent Interaction*, pp. 622-628, 2005.

[15] C. Elliott, "The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System," PhD dissertation, Northwestern Univ. The Inst. for the Learning Sciences, Technical Report No. 32, 1992.

[16] A.C. Boucouvalas, "Real Time Text-to-Emotion Engine for Expressive Internet Communications," *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*, pp. 306-318, Ios Press, 2003.

[17] P. Subasic and A. Huettner, "Affect Analysis of Text Using Fuzzy Semantic Typing," *IEEE Trans. Fuzzy Systems*, vol 9, pp. 483-496, 2001.

[18] J. Olveres, M. Billinghurst, J. Savage, and A. Holden, "Intelligent, Expressive Avatars," *Proc. First Workshop Embodied Conversational Characters*, pp. 47-55, 1998.

[19] L. Devillers, I. Vasilescu, and L. Lamel, "Annotation and Detection of Emotion in a Task-Oriented Human-Human Dialog Corpus," *Proc. ISLE Workshop Dialogue Tagging for Multi-Modal Human Computer Interaction*, 2002.

[20] J. Tao and T. Tan, "Emotional Chinese Talking Head System," *Proc. Sixth Int'l Conf. Multimodal Interfaces*, pp. 273-280, 2004.

[21] A. Wensel and S.O. Sood, "VIBES: Visualizing Changing Emotional States in Personal Stories," *Proc. Second ACM Workshop Story Representation, Mechanism and Context*, pp. 49-56, 2008.

[22] H. Liu, H. Lieberman, and T. Selker, "A Model of Textual Affect Sensing Using Real-World Knowledge," *Proc. Int'l Conf. Intelligent User Interfaces*, pp. 125-132, 2003.

[23] Y.S. Seol, D.J. Kim, and H.W. Kim, "Emotion Recognition from Text Using Knowledge-Based ANN," *Proc. Int'l Technical Conf. Circuits/Systems, Computers and Comm.*, pp. 1569-1572, 2008.

[24] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect Analysis Model: Novel Rule-Mased Approach to Affect Sensing from Text," *Int'l J. Natural Language Eng.*, vol. 17, no. 1, pp. 95-13, 2011.

[25] C.O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction," *Proc. Human Language Technology Conf., Conf. Empirical Methods in Natural Language Processing*, pp. 579-586, 2005.

[26] S. Aman and S. Szpakowicz, "Identifying Expressions of Emotion in Text," *Proc. 10th Int'l Conf. Text, Speech and Dialogue*, pp. 196-205, 2007.

[27] P. Katz, M. Singleton, and R. Wicentowski, "Swat-mp: The SemEval-2007 Systems for Task 5 and Task 14," *Proc. Fourth Int'l Workshop Semantic Evaluations*, pp. 308-313, 2007.

[28] C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective Text," *Proc. Fourth Int'l Workshop Semantic Evaluations*, pp. 70-74, 2007.

[29] G. Mishne, "Experiments with Mood Classification in Blog Posts," *Proc. First Workshop Stylistic Analysis of Text for Information Access*, 2005.

[30] G. Leshed and J. Kaye, "Understanding How Bloggers Feel: Recognizing Affect in Blog Posts," *Extended Abstracts Int'l Conf. Computer-Human Interaction*, pp. 1019-1024, 2006.

[31] R. Mihalcea and H. Liu, "A Corpus-Based Approach to Finding Happiness," *Proc. AAAI Spring Symp. Computational Approaches to Weblogs*, pp. 139-144, 2006.

[32] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," *Data-Centric Systems and Applications*, Springer, first ed., 2007.

[33] A. Boucouvalas and X. Zhe, "Text-to-Emotion Engine for Real Time Internet Communication," *Proc. IEEE, IET Int'l Symp. Comm. Systems, Networks and Digital Signal Processing*, pp. 164-168 2002.

[34] M. Mulder, A. Nijholt, M.d. Uyl, and P. Terpstra, "A Lexical Grammatical Implementation of Affect," *Proc. Seventh Int'l Conf. Text, Speech and Dialogue*, pp. 171-178, 2004.

[35] K. Moilanen and S. Pulman, "Sentiment Composition," *Proc. Int'l Conf. Recent Advances in Natural Language Processing*, pp. 378-382, 2007.

[36] L. Wittgenstein, *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*, compiled from notes taken by Y. Smythies, R. Rees, and J. Taylor, C. Barnet, ed. Univ. of California Press, 1967.

[37] J. Gajadhar and J. Green, "An Analysis of Nonverbal Communication in an Online Chat Group," The Open Polytechnic of New Zealand, working paper, 2003.

[38] Synesketch, Creative Review, Nov. 2008.

[39] O.d. Pos and P. Green-Armytage, "Facial Expressions, Colours and Basic Emotions," *Colour: Design & Creativity*, Soc. of Dyers and Colorists, 2007, http://aic-colour-journal.org/index.php/JAIC/article/viewFile/77/71, Oct. 2012.

[40] F. Radulovic and N. Milikic, "Smiley Ontology," *Proc. SNI*, 2009.

[41] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.

[42] Z.J. Chuang and C.H. Wu, "Multi-Modal Emotion Recognition from Speech and Text," *Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 45-62, 2004.

[43] M. Purver and S. Battersby, "Experimenting with Distant Supervision for Emotion Classification," *Proc. 13th Conf. European Chapter Assoc. for Computational Linguistics*, pp. 482-491, 2012.

[44] Z. Yuan and M. Purver, "Predicting Emotion Labels for Chinese Microblog Texts," *Proc. First Int'l Workshop Sentiment Discovery from Affective Data*, pp. 40-47, 2012.

[45] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," master's thesis, Stanford Univ., 2009.

[46] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Proc. Seventh Conf. Int'l Language Resources and Evaluation*, 2010.

[47] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text," *J. Am. Soc. for Information Science and Technology*, vol. 61, no. 12, pp. 2544-2558, 2010.

[48] H. Liu and P. Singh, "ConceptNet: A Practical Commonsense Reasoning Tool-Kit," *BT Technology J.*, vol. 22, no. 4, pp. 211-226, 2004.

[49] R.A. Calix, S.A. Mallepudi, B. Chen, and G.M. Knapp, "Emotion Recognition in Text for 3-D Facial Expression Rendering," *IEEE Trans. Multimedia*, vol. 12, no. 6, 2010.

[50] C.O. Alm, *Affect in Text and Speech*. VDM Verlag, 2009.

[51] M. Lehrman, C.O. Alm, and R.A. Proano, "Detecting Distressed vs. Non-Distressed Affect State in Short Forum Texts," *Proc. Workshop Language in Social Media at the Conf. North Am. Chapter Assoc. for Computational Linguistics-Human Language Technologies*, pp. 9-18, 2012.

[52] F. Keshtkar, "A Computational Approach to the Analysis and Generation of Emotion in Text," PhD dissertation, Univ. of Ottowa, Canada, 2011.

[53] V. Francisco and P. Gervas, "EmoTag: An Approach to Automated Markup of Emotions in Texts in Computational Intelligence," *Computational Intelligence*, 2012.

[54] U. Krcadinac, J. Jovanovic, and V. Devedzic, "Visualizing the Affective Structure of Students Interaction," *Proc. Fifth Int'l Conf. Hybrid Learning*, pp. 23-34, 2012.

[55] D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, section 3.2.1. Addison-Wesley, 1969, third ed., 1997.

[56] A.P. Field, *Discovering Statistics Using SPSS: And Sex and Drugs and Rock'n'Roll*, third ed. Sage, 2009.

[57] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 79-86, 2002.

[58] S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, vol. 53, no. 9, pp. 1375-1388, 2007.

Uros Krcadinac is a research assistant and a PhD candidate in the Department of Software Engineering, University of Belgrade, Serbia, where he also received the BS degree in informatics and software engineering in 2008. His research interests include affective computing, information and data visualization, text mining, data storytelling, and new media art. He is a member of the GOOD OLD AI Research Network and the Metacreation Lab. His homepage is http://krcadinac.com/.

Philippe Pasquier received the PhD inartificial intelligence from Laval University. He is an assistant professor and the Graduate Program Chair at the School of Interactive Arts and Technology, and an adjunct professor in the Cognitive Science Program at Simon Fraser University. He heads the Metacreation, Agents, and Multiagent Systems laboratory, which focuses on the development of models and tools for endowing machines with autonomous, intelligent, or creative behavior. He is a member of AAAI, the ACM, and the Cognitive Science Society. Contact him at pasquier@sfu.ca.

Jelena Jovanovic is an associate professor in the Department of Software Engineering, University of Belgrade, Serbia. Her research interests are in the areas of knowledge representation and semantic technologies, and their application in the domain of technology enhanced learning and knowledge management. She has participated in a number of projects and published numerous refereed papers in these research areas. She has also been teaching software engineering, semantic technologies, and related AI technologies both at the undergraduate and postgraduate level. She can be reached at http://jelenajovanovic.net.

Vladan Devedzic is is a professor in the Department of Software Engineering, University of Belgrade, Serbia. His research interests include knowledge modeling, ontologies, Semantic Web, intelligent reasoning techniques, software engineering, and application of artificial intelligence to education and healthcare. He is the founder and the chair of the GOOD OLD AI Research Network. His homepage is http://devedzic.fon.rs/