



UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

Automatic detection of maintenance requests: Comparison of Human Manual Annotation and Sentiment Analysis techniques

This is the peer reviewed version of the following article:

Original

Automatic detection of maintenance requests: Comparison of Human Manual Annotation and Sentiment Analysis techniques / D'Orazio, M.; Di Giuseppe, E.; Bernardini, G.. - In: AUTOMATION IN CONSTRUCTION. - ISSN 0926-5805. - ELETTRONICO. - 134:(2022). [10.1016/j.autcon.2021.104068]

Availability:

This version is available at: 11566/295881 since: 2024-04-22T12:29:21Z

Publisher:

Published

DOI:10.1016/j.autcon.2021.104068

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.

This item was downloaded from IRIS Università Politecnica delle Marche (<https://iris.univpm.it>). When citing, please refer to the published version.

(Article begins on next page)

1 **Towards a better automatic detection of maintenance requests: a comparison of Human** 2 **Manual Annotation and different sentiment analysis techniques**

3 Marco D'Orazio¹, Elisa Di Giuseppe¹, Gabriele Bernardini^{1,*}

4
5 1-DICEA Department, Università Politecnica delle Marche, via Brecce Bianche, 6013 Ancona (Italy);
6 corresponding author*: g.bernardini@staff.univpm.it

7 8 9 **Abstract**

10 In the building management process, the collection of end-users' maintenance request is a rich source
11 of information to evaluate occupants' satisfaction and building systems. Computerized Maintenance
12 Management Systems typically collect non-standardized data, difficult to be analysed. Text mining
13 methodologies can help to extract information from end-users' maintenance requests and support
14 priority assignment of decisions. Sentiment Analysis (SA) can be applied to this end, but complexities
15 due to words/sentences orientations/polarities and domains/contexts can reduce their effectiveness.
16 Human Manual Annotation (HMA) could better support this process. This study compares the ability of
17 different SA techniques and HMA to automatically define a maintenance severity ranking. About 12.000
18 requests were collected for 34 months in 23 buildings of a University Campus. Results show that,
19 differently from SA, HMA takes advantages of technical words recognition, providing a better
20 assessment of requests severity and representing the first step for future lexicon development.

21 22 **Keywords**

23 Facility management, maintenance, facility management, human manual annotation, Sentiment
24 analysis

25 26 **1. Introduction**

27 In line with the advancement of technology, building management has entered into a digital era [1–4].
28 In addition to data mining, text mining has become a fundamental tool to discover hidden knowledge
29 from massive and complex data stored in databases or other information repositories, including
30 patterns, correlations, relationships, and anomalies [5]. Automatic systems for data analysis in the
31 contexts of building constructions can take advantage of such techniques to improve the building
32 management quality, decrease the maintenance costs, timely react to building faults or other critical
33 conditions under different circumstances (including emergencies), and thus increase the end-users'
34 satisfaction [4,6–8].

35 Sentiment analysis recently received particular attention in the field of facility management, due to the
36 importance of end-user perceptions and opinions about building Operation and Maintenance (O&M)
37 activities. These methodologies can help to collect information about the status of building systems,
38 directly from end-users perceptions [9], to improve dynamically preventive maintenance strategies [2].
39 Sentiment analysis [10] is the computational study of people's opinions, sentiments, emotions, and
40 attitudes [11,12], often employed to extract opinion polarity and degree [13] from different sources
41 [14,15]. The rapid growth of sentiment analysis application coincides with the growth of reviews, forum,

42 discussions, blogs, and microblogs on social media, and the growth of a huge volume of opinion data
43 recorded in digital forms [11].

44 Consequently, the volume and diversity of research articles applying sentiment analysis are expanding
45 rapidly. However, sentiment analysis is a complex task [10]. It is well known the most important
46 indicators of sentiments are sentiment words, also called “opinion” words [11]. Moreover, there are also
47 phrases and idioms expressing sentiments. A list of such words and phrases is called a sentiment
48 lexicon (or opinion lexicon). Over the years, researchers have designed numerous algorithms to compile
49 such lexicons. Although sentiment words and phrases are important, they cannot provide accurate
50 sentiment analysis on their own. A positive or negative sentiment word may have opposite orientations
51 or polarities in different application domains or sentence contexts. A sentence containing sentiment
52 words may even not express any sentiment. Sarcastic sentences with or without sentiment words are
53 hard to deal with. Many sentences without sentiment words can even imply positive or negative
54 sentiments or opinions of their authors [11]. Finally, many words or sentences may have opposite
55 orientations or polarities in different application domains [10].

56 Recently, sentiment analysis methodologies have been also applied to analyze several aspects of the
57 building management process. Marzouk and Enaba [16] developed a Dynamic Text Analytics for
58 Contract and Correspondence (DTA-CC) model to monitor correspondence sentiment and
59 communication nature. Text mining techniques [17] were applied to identify the major treated topics
60 related to the energy use and management of buildings and to collect information about energy policy
61 preferences and concerns. Loureiro and Alló [18] employed a Natural Language Processing (NLP)
62 tools, based on the lexicon developed by the National Research Center Canada (NRC), denoted as
63 EmoLex [19]. The NRC Emotion Lexicon contains a list of English words and their associations with
64 eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two main
65 sentiments (negative and positive) [20]. Sun et al. analyzed microblog posts to derive information about
66 opinions on operational aspects such as energy policies [21]. Positive and negative words are quantified
67 basing on the China HowNet Thesaurus. Liu and Hu performed sentiment analysis of the public
68 attention status and changing trends toward green buildings, based on Ekman’s six basic universal
69 emotions [22].

70 More recently, natural language processing models were applied to the facility management of
71 buildings, collecting sentiments and opinions from end-users, to improve the building operability and
72 the cost of the management process [23,24]. Bortolini and Forcada developed a methodology, based
73 on the TF analysis of words expressing the severity degree, to determine the typical problems that end-
74 users complain about the building systems and their perceived severity [9]. Gunay et al. analyzed
75 operators’ work order descriptions in Computerized Maintenance Management Systems (CMMS),
76 extracting information about failure patterns in building systems and components [25]. The results
77 provide insights into equipment breakdown of failure events, top system and component-level failure
78 modes, and their occurrence frequencies. Bouabdallaoui et al. proposed a machine-learning algorithm
79 based on Natural Language Processing (NLP) to manage day-to-day maintenance activities [26].
80 Sexton et al. compared NLP methodologies to extract keywords from maintenance Work orders [27].
81 Bardhan et al. employed two emotion lexicon databases, the Ho-Liu database [11] and the NRC

82 emotion lexicon from semi-structured interviews and focus group discussions regarding housing
83 management in India [28]. The author justifies the choice of two lexicons arguing that the Ho-Liu lexicon
84 is a tool to understand the general sentiments of the documents in a binary fashion, considering only
85 positive or negative sentiment as categorized in [11], while the NRC lexicon enables the classification
86 of the sentiments into discreet emotions [19]. Sun et al. calculated sentiment value on energy price
87 policies basing on polarity and intensity of sentiment words, based on the China HowNet Thesaurus
88 [21]. The authors adopted a sentence pattern based on sentiment words, privative, degree words and
89 rhetorical question.

90 Several general-purpose subjectivities, sentiment, and emotion lexicons have been realized and are
91 publicly available [11,19,29–32], but the accuracy of proposed methodologies and lexicons should be
92 properly evaluated when applied to specific domains or to extract specific sentiments related to some
93 aspect of the sentence. Sharma and Dutta showed that sentiment lexicons are convenient since they
94 are much faster and less computationally intensive compared to Machine Learning (ML) methods [33].
95 Moreover, ML models don't generalize well and perform poorly when used in a different domain.

96 Several studies have been performed to check the concordance of different lexicons in different
97 domains. Some of them studied the problem of polarity or orientation consistency checking among
98 sentiment lexicons or dictionaries [34,35]. Schmidt and Burghardt evaluated the performance of
99 different German sentiment lexicons and processing configurations like *lemmatization*, the extension of
100 lexicons with historical linguistic variants and stop words elimination, in order to explore the influence
101 of these parameters and to find best practices for a specific domain of application [36]. A comparative
102 study on sentiment analysis approaches and methods analyzed machine learning, rule-based and
103 lexicon-based methods, together with different machine learning methods (as SVM, N-gram SA, NB,
104 ME, KNN methods and multilingual approach) [37]. Based on a state of the art, the author showed that
105 the accuracy of different methods could range from 66% to 95.5%. To investigate the relationship
106 between sentiment analysis approaches and social context, Sánchez-Rada and Iglesias proposed a
107 framework, also evaluating the performance of different techniques applied in different contexts [38].

108 Various combinations of existing lexicons and NLP tools have been evaluated against a human-
109 annotated subsample [39], which serves as a gold standard. In fact, Human manual Annotation (HMA)
110 techniques still seem to better retrieve the presence of particular terms (i.e. technical words) having a
111 paramount role depending on the domain and context of the application. Cambria et al. described
112 several comparative works, based on human annotation approaches (Best-Worst, MaxDiff) [10]. Borg
113 and Boldt investigated sentiment analysis in customer support for a large Swedish Telecom corporation,
114 comparing VADER Valence-Aware sentiment lexicon with annotations of human experts [40]. The best
115 performing configuration accomplished an accuracy of 70%.

116 However, despite a significant amount of research, challenging problems remain. In this context, a
117 general and effective method for discovering and determining domain and context-dependent
118 sentiments is still lacking [41]. It is hence necessary to preliminarily check the accuracy of proposed
119 methodologies when applied to each specific domain to extract information about specific aspects.
120 Then, a wide comparison between sentiment analysis techniques and HMA methods should be
121 provided to better assess differences and similarities between them, especially when moving towards

122 the automatic detection of the priority order in maintenance requests, which is a paramount element to
123 support O&M [42]. Indeed, the immediate and automatic detection of the severity (importance and
124 urgency) of any maintenance request, through text mining methodologies, could reduce the risks
125 associated with late interventions and improve preventive maintenance strategies, providing useful
126 information to change on-the-fly planned activities and reducing buildings' O&M costs.

127 Given the context of buildings maintenance, this study tries to compare different sentiment lexicons and
128 an HMA method (developed in this work) to assess the severity of maintenance's requests depending
129 on the end-users' non-standardized communications. Eleven polarity-based and valence-based
130 lexicons were compared with a text mining approach based on the recognition of words expressing
131 different severity levels (SSA) and with a human annotation scheme (HMA) based on BWS (Best-Worst)
132 methodology. The analyzed dataset includes the maintenance requests collected from January 2018
133 to October 2020 by the end-users of a University organization comprising 23 buildings.

134

135 **2. Related works on sentiment analysis lexicons**

136 Under the umbrella of sentiment analysis, there are different tasks and methodologies. Sentiment
137 analysis research can be carried out at different levels: document, sentence, and aspect [11], obtaining
138 different results. At the document level, it is possible to classify whether a whole opinion document
139 expresses a positive or negative sentiment. At the sentence level, it is possible to determine whether
140 each sentence expresses a positive, negative, or neutral opinion. However, a sentence could even
141 comprise general positive opinions but not related to specific aspects, services or products. Instead of
142 looking at language units (documents, paragraphs, sentences, clauses, or phrases), aspect-level
143 analysis directly looks at opinion and its target (called opinion target) [11]. Based on this level of
144 analysis, a summary of opinions about entities and their aspects can be produced. Several general
145 lexicons have been realized and are available to perform these tasks, i.e. General inquirer lexicon, HU-
146 LIU lexicon [11], MPQA subjectivity lexicon [29], SentiWordNet [30,31], EmoLex lexicon [19,32].

147 Borg and Boldt proposed a classification of lexicons into two main groups [40]: Semantic orientation
148 (polarity-based) lexicons; Sentiment intensity (valence-based) lexicons. Table 1 reports the main
149 characteristics of several publicly available lexicons and the tool where they are implemented (R
150 statistics - rel. 4.0 - packages).

151 The first group comprises lexicons containing a list of lexical features (e.g. words) which are generally
152 labelled according to their semantic orientation as either positive or negative. The oldest semantic
153 orientation lexicons are part of proprietary text-analysis software, such as LIWC and GI (General
154 Inquirer). But also public polarity-based lexicons are available. [43] maintains a publicly available lexicon
155 of nearly 6,800 words (2,006 with positive semantic orientation, and 4,783 with negative semantic
156 orientation). WordNet [31] is a well-known English lexical database in which words are clustered into
157 groups of synonyms known as synsets. Other polarity-based lexicons, described in [10] are
158 SentiWordNet (WordNet improvement), SO-CAL, AFINN, QDAP, and specific domain lexicons such as
159 Henry Financial and Loughran-McDonald.

160

| Lexicon | Type | General information | Ref. | Tool |
|---------|------|---------------------|------|------|
|---------|------|---------------------|------|------|

| | | | | |
|----------------|---------------------------------|--|------|-----------------------|
| GI | Polarity-based | List of positive and negative words according to the psychological Harvard-IV dictionary as used in the General Inquirer software. | [48] | SentimentAnalysis (R) |
| HU-LIU | Polarity-based | General-purpose English sentiment lexicon that categorizes positive (1) and negative (-1) words. | [43] | Sentimentr (R) |
| NRC | Polarity-based | List of positive (1) and negative (-1) words (3241 Negative and 2227 positive words) | [32] | Sentimentr (R) |
| HE | Polarity-based | List of positive and negative words according to the Henry's finance dictionary (53 positive, 44 negative) | [49] | SentimentAnalysis (R) |
| LM | Polarity-based | List of positive, negative and uncertainty words according to the Loughran-McDonald finance-specific dictionary (185 positive, 885 negative) | [50] | SentimentAnalysis (R) |
| QDAP | Polarity-based | List of polarity words part of qdap package. 2952 negative words, 1280 positive words | [51] | SentimentAnalysis (R) |
| AFINN | Valence-based | List of English terms manually rated for valence with an integer between -5 (negative) and +5 (positive) | [52] | Syuzhet (R) |
| SENTIWORDNET | Valence-based | Lexicon in which each WORDNET synset is associated to three numerical scores, describing how objective, positive, and negative the terms contained in the synset are. Each of the three scores ranges from 0 to 1 and their sum is 1 | [53] | Sentimentr (R) |
| SenticNet | Valence-based | List of positive and negative word associated with a numerical score ranging from -1 to 1 (23626 words) | [54] | Sentimentr (R) |
| Jockers | Valence-based | List of positive and negative words associated with a numerical score ranging from -1 to 1. (10738 words) | [55] | Sentimentr (R) |
| Jockers-Rinker | Valence-based | Combined and augmented version of Jockers & Rinker's augmented Hu-Liu lexicon, containing a list of positive and negative words associated with a numerical score ranging from -1 to 1. (10738 words) | [55] | Sentimentr (R) |
| VADER | Valence-based and lexical rules | List of 7500 lexical features with valence scores expressing sentiment intensity ranging from -4 to 4 | [44] | Vader (R) |

Table 1 Several publicly available lexicons organized according to Borg and Boldt's classification (second column) [40].

161
162

163 The second group comprises lexicons useful to determine not just the binary polarity (positive versus
164 negative), but also the strength of the sentiment expressed in text. Thus, sentiment intensity lexicons
165 can recognize the strength of sentiment. Sentiment intensity lexicons have been further improved with
166 disambiguation processes and mixing lexical features with rules that embody grammatical and
167 syntactical conventions used by humans when expressing or emphasizing sentiment intensity [44].
168 VADER (Valence Aware Dictionary for sEntiment Reasoning) is a sentiment intensity lexicon that
169 combines quantitative and qualitative features. The Affective Norms for English Words (ANEW) lexicon
170 provides a set of normative emotional ratings for 1,034 English words [45]. ANEW words have an
171 associated sentiment valence ranging from 1-9. SentiWordNet (SWN) is an extension of WordNet in
172 which 147,306 synsets are annotated with three numerical scores relating to positivity, negativity, and
173 neutrality [30]. SentiWords, an high coverage lexicon for sentiment analysis based on SentiWordNet
174 [46]. SenticNet is a publicly available semantic and affective resource for concept-level opinion and
175 sentiment analysis [10]. The SenticNet lexicon consists of 14,244 common sense concepts such as
176 wrath, adoration, woe, and admiration with information associated with (among other things) the
177 concept's sentiment polarity, a numeric value on a continuous scale ranging from -1 to 1. More recently
178 also emotion lexicons were introduced. NRC Emolex (also called NRC Word-Emotion Association
179 Lexicon, described in [19]) classifies sentiment by mapping a large list of emotions into eight basic

180 groups of emotions: trust (acceptance, admiration, like); fear (fear); surprise (uncertainty, amazement),
181 sadness (sadness), disgust (dislike, hate, dis- appointment, indifference) anger (anger), anticipation
182 (anticipation and vigilance) and joy (calmness, joy) into a four-point scale in addition to the positive and
183 negative words [20]. Gatti et al. introduced other available emotion lexica: NRC Hashtag, NRC Affect,
184 WordNet-Affect (wordnet extension); AffectNet; Fuzzy Affect Lexicon; Emolex; Affect; DepecheMood
185 ++ [46]. DepecheMood++, also called DM++, is a bi-lingual lexicon (English- Italian) improvement of
186 DepecheMood, developed in [47]. DM++ and has been compared with Hu-Liu, MPQA, NRC-Emolex,
187 SentiWordNet lexicons in the task of emotion intensity prediction.

188

189 **3. Methodology**

190 3.1. Collecting end-user maintenance requests and generating the work orders (WOs)

191 This work is based on the evaluation of end-users' requests concerning the maintenance interventions
192 on the building stock of the University "Politecnica delle Marche" (UNIVPM) located in Ancona, Italy.
193 UNIVPM building stock comprises 23 buildings and hosts a population of about 16.000 students and
194 1000 workers. The facility management activity of UNIVPM is performed through a CMMS, by a general
195 contractor (ANTAS). The contractor grants both the predictive maintenance service (e.g. components'
196 replacement before their expected end-of-life) and the on-demand service (e.g. components' repair or
197 replacement after faults complained by end-users through e-mails).

198 End-user's maintenance requests are short texts exchanged by e-mail and processed by contractor
199 technicians. In the process, each end-user's request is translated into a Work Order (WO) by the
200 technicians, by joining the text of the mail with technical information (e.g. system typology by class and
201 subclass, date, ID) after a preliminary check to evaluate the consistency of the request. WOs then
202 comprise a mix of end-user's personal perceptions and technical information. During the busiest days,
203 the technical staff receive at least 20-30 different WOs.

204 The analyzed dataset comprises communications (WO) about anomalies and faults in the buildings'
205 components and systems and related maintenance interventions, collected from January 2018 to
206 October 2020, hence also during the COVID-19 emergency. The dataset comprises 7 WO categories:
207 electrical (lighting, power systems, LAN and WLAN connection), building components (walls doors,
208 windows, floors, stairs); HVAC (heating, ventilation and cooling units and pipes); plumbing (plumbing
209 and sanitary systems); fire (fixed and moveable devices); dialer alarm (alarm systems); elevator (cabins,
210 motors).

211

212 3.2. WO's Text mining

213 After a preliminary evaluation of the metric of the sentences by category, and considering that the WOs
214 corpus is a single document including requests comprising 10274 paragraphs and 11.449 sentences,
215 a TF (Term Frequency) analysis [56,57] has been performed to extract information about the most
216 frequent aspects of intervention requests (nouns), the actions required (verbs) to solve the problem and
217 the characteristic of the problem (adjectives and adverbs). Texts were preliminarily treated to remove
218 stop-words, punctuation, symbols, etc... A stemming process to reduce inflected and derived words to
219 their root form have been performed. TF calculates the frequency of a word appearing in the document.

220 Metric and TF analysis have been performed through R statistics software (ver. 4.0) and the “quanteda”,
221 “tm” and “SnowballC” text mining packages. To evaluate the association between the nouns used to
222 describe the problems, a “word association” analysis has been performed on the most frequent words,
223 and the Jaccard similarity score has been calculated. Jaccard similarity ranges from 0 to 1 and refers
224 to the number of common words overall words of the end-user maintenance corpus. Moreover, a
225 “classical multidimensional scaling analysis” has been performed to visualize in a 2 N-dimension space
226 the level of similarity of the end-users requests of the dataset. Jaccard similarity has been used to
227 represent the distance among individuals. Indeed, Jaccard similarity coefficient is used for measuring
228 the similarity and diversity of sample sets and it is defined as the size of the intersection divided by the
229 size of the union of the sample sets. Finally, a Co-occurrence network comprising nouns, verbs,
230 adjectives, and adverbs has been realized to visualize the potential relationships between aspects and
231 characteristics of the intervention requests and actions required to solve the problem. Co-occurrence
232 network of terms is based on their paired presence within a specified unit of text (sentence). Networks
233 are generated by connecting pairs of terms using a set of criteria defining co-occurrence. “Word
234 association”, “Classical multidimensional scaling maps” and “co-occurrence network” have been
235 realized through KHcoder text mining code [58,59].

236

237 3.3. Human manual annotation and semi-automatic human annotation

238 To define a gold standard useful to check the validity of the sentiment analysis approach based on
239 lexicons, a human annotation scheme (HMA) based on the best-worst scaling (BWS) approach [60]
240 has been performed. The best-worst scaling technique (BWS) is a variant of comparative annotations
241 proposed in [61]. BWS addresses the limitations of traditional rating scales [62] working on n-tuples.
242 Annotators are presented with n items at a time (an n-tuple, where $n > 1$, and typically $n = 4$). They are
243 asked which item is the best (highest in terms of the property of interest) and which is the worst (lowest
244 in terms of the property of interest). When working on 4-tuples, best–worst annotations are particularly
245 efficient because by answering these two questions, the results for five out of six item–item pair-wise
246 comparisons become known.

247 In this work, annotators were presented with several 4-tuples and asked to select the most positive and
248 the most negative. A random subset of sentences has been extracted from the dataset, respecting the
249 proportion of sentence by category type. 150 distinct 4-tuples were randomly generated through the
250 “*bwstuples*” python script (<http://valeribasile.github.io/>), in such a manner that each term was seen in
251 five different 4-tuples. Each 4-tuple was annotated by 13 experts with different expertise. Three groups
252 were defined depending on their expertise in the building O&M field: HE (High Expertise) group, made
253 by 5 annotators with at least 10 years of expertise in the field; NE (Normal Expertise) group, made by
254 3 annotators with at least 5 years of expertise in the field; LE (Limited Expertise) group, made by 5
255 annotators with at least 2 years of expertise in the field. The score is given by the number of times an
256 item chosen as BEST – WORST divided by the number of times an item appears [61,62]. The final
257 score for each WO is the mean of scores given by each annotator.

258 Calculated Human Manual Annotation (HMA) scores have been then translated into a three-level scale
259 (Negative, Neutral, Positive) assuming a “Negative” polarity for scores in the range “-1:-0.33”, a

260 “Positive” polarity for scores in the range “0.33:1” and “Neutral” in the range “-0.33:0.33” scores. The
261 three levels are then characterized by the same size. A polarity annotation contingency table has been
262 plotted to evaluate the agreement of all annotators and the Krippendorff’s α coefficient has been
263 calculated [39].

264 An alternative approach based on [9] has been introduced to check the possibility of a semi-automated
265 annotation approach (SSA). The SSA is based on the detection of the most frequent words associated
266 with high, medium, and low severity issues. According to [9] we considered three levels of severity (low,
267 medium and high). “High”, “medium” and “low” scores attributed through SSA correspond to HMA
268 “Negative”, “Neutral” and “Positive” levels. High severity words are typically used when an immediate
269 repair or action is required (e.g. urgent, safety, emergency, alarm, fire). Low severity words are typically
270 used when a repair or action can be slightly postponed (e.g. adjust, install, verify, check, replace, clean,
271 paint). Low severity words are used to communicate low-impact events without requiring urgent or
272 planned actions. The list of “high severity” and “low severity” words has been manually derived by three
273 experts from the results of the TF analysis, selecting the terms expressing high severity or low severity
274 where annotators agree. According to [9] we assumed mean severity words as the words not labelled.
275 Then the presence of the most frequent words related to high, medium, and low severity was checked
276 for each sentence. Each sentence (representing a WO) was labelled as “high”, “medium”, “low” severity
277 according to the presence of at least one of these words. Labelling has been performed employing R
278 statistics software (rel. 4.0) and related text mining packages.

279

280 3.4. Sentiment and emotion analysis

281 To understand the ability of polarity-based and valence-based lexicons to detect the severity of end-
282 user maintenance requests, we choose 11 publicly available polarity-based lexicons (GI [48], AFINN
283 [51], Hu-Liu [43], SentiwordNet [53], NRC [63], Senticnet [54], Jockers [55], Jockers-Rinker [55], HE
284 [49], LM [50], QDAP [51]) and 1 valence-based lexicons (VADER [44]).

285 The analysis has been performed through R statistics (rel. 4.0), and the following packages:
286 “Sentimentr” [64], “Syuzhet”, “SentimentAnalysis”, “Lexicon”, and “Vader” [47]. “Sentimentr” is the
287 bridge towards the lexicons: Hu-Liu, NRC, Sentiword, Senticnet, Jockers and Jockers-Rinter. Through
288 “Syuzhet”, lexicon AFINN is available and through SentimentAnalysis GI, HE, LM and QDAP are
289 available.

290 The equation used by the “Sentimentr” algorithm to assign scores utilizes lexicons to tag polarized
291 words. Each paragraph ($p_i = \{s_1, s_2, \dots, s_n\}$) composed of sentences, is broken into element sentences
292 ($s_{i,j} = \{w_1, w_2, \dots, w_n\}$) where w are the words within sentences. Each sentence (s_j) is broken into an
293 ordered bag of words. Punctuation is removed except for pause punctuations (commas, colons,
294 semicolons) which are considered a word within the sentence. The words in each sentence ($w_{i,j,k}$) are
295 searched and compared to the chosen dictionary of the lexicon package. Positive ($w_{i,j,k+}$) and negative
296 ($w_{i,j,k-}$) words are tagged with a +1 and -1 respectively (or other positive/negative weightings
297 depending on the sentiment dictionary). Polarized words form a polar cluster ($c_{i,j,l}$) which is a subset
298 of the sentence where j and l are the words before and after positive or negative polarized words. After
299 preliminary tests, the polarized context cluster ($c_{i,j,l}$) of words is pulled from around the polarized word

300 (p**w) and 4 words before and 2 words after (p**w) were considered as valence shifters. The words in
301 this polarized context cluster are tagged as neutral (wi, j, k0), negator (wi, j, kn), amplifier [intensifier]
302 (wi, j, ka), or de-amplifier [downtoner] (wi, j, kd). Each polarized word has been weighted (w) assuming
303 the "polarity_dt" argument = 0.8 and then further weighted by the function and number of the valence
304 shifters directly surrounding the positive or negative word (p**w). Valence shifters are: amplifiers/de-
305 amplifiers (i.e double negations shifting the polarity); adversative conjunctions (i.e., 'but', 'however', and
306 'although') before and after the polarized word. Adversative conjunction makes the next clause of
307 greater values while lowering the value placed on the prior clause. Finally, the weighted context clusters
308 of each sentence are summed and divided by the square of the word count yielding an unbounded
309 polarity score for each sentence. Considering that the text of each WO comprises one or more
310 sentences, the WO score has been calculated by grouping sentence score by the identifier code,
311 obtaining a mean score and relative standard deviation in case of multiple sentences in the same text.
312 Syuzhet package is the key access to the AFINN dictionary, where each word is associated with a
313 polarity score (-1;1). Each sentence has been broken into an ordered bag of words. Numbers,
314 punctuation and extra-spaces have been removed and the words in each sentence are searched and
315 compared to the chosen dictionary of the lexicon package. Sentence score has been calculated by
316 "syuzhet" package as the sum of scores associated with each polarized word.
317 "SentimentAnalysis" package is the key access to GI, HE, LM and QDAP polarity-based lexicons.
318 The package functions calculate the sentiment score for each sentence according to the following
319 approach: number of positive words minus the number of negative words in respect to the whole number
320 of words. As previously described each sentence has been broken into an ordered bag of words,
321 numbers, punctuations and extra-spaces have been removed and the words were compared with GI,
322 HE, LM and QDAP dictionaries. Sentence score has been calculated by "SentimentAnalysis" package
323 as the difference between the sum of positive and negative words in respect to the polarized words of
324 the sentence.
325 "Vader" package has been used to perform sentiment analysis through VADER [44] (Valence Aware
326 Dictionary for sEntiment Reasoning). VADER combines lexical features with consideration for five
327 generalizable rules that embody grammatical and syntactical conventions that humans use when
328 expressing or emphasizing sentiment intensity. Incorporating these heuristics improves the accuracy of
329 the sentiment analysis engine across several domain contexts [44]. VADER aggregate sentiment
330 scores from individual words into sentence scores [40]. The methodology comprises the calculation of
331 four "sentiment" scores (positive, negative, neutral, compound). The compound score is a synthetic
332 sentence score computed by summing the valence scores of each word in the lexicon, adjusted
333 according to the lexical rules, and then normalized to be between -1 (most extreme negative) and +1
334 (most extreme positive) [40,44].

335 336 3.5. Comparison methodology

337 HMA has been assumed as the gold standard [36] to measure the ability of the other methods to
338 correctly evaluate the WO's severity. HMA results were expressed both on a numeric scale and on a

339 three-level scale (negative, neutral, positive). The score conversion into a three-level scale is justified
 340 by the necessity to compare HMA with methods characterized only by a three-level scale (i.e SSA) [9].
 341 Firstly, SSA results have been compared with the HMA according to the three ranking scales previously
 342 described. Precision, Recall and F1 measure [65,66] have been used to compare results by groups
 343 (Table 2). Recall is the ratio of the number of elements correctly classified to the number of known
 344 elements in each class. Precision is the ratio of the number of elements correctly classified to the total
 345 predicted in each class. F1 measure is the harmonic mean between both precision and recall. In detail,
 346 the precision of the negative class is computed as: $P(\text{neg})= i/(c + f + i)$; its recall, as: $R(\text{neg})= i/(g + h +$
 347 $i)$; and $F1(\text{neg})= [2P(\text{neg}) * R(\text{neg})] / [P(\text{neg})+R(\text{neg})]$.

| | | SSA | | |
|-----|---------|----------|---------|----------|
| | | Positive | Neutral | Negative |
| HMA | Best | a | b | c |
| | Neutral | d | e | f |
| | Worst | g | h | i |

348 *Table 2 Confusion matrix for experiments with three classes [66].*

349 Then, comparisons between different lexicons and between HMA and lexicons have been performed
 350 through a statistical analysis based on the calculation of the Spearman correlation coefficient, after a
 351 normalization process, to obtain data characterized by mean=0 and sd=1. Spearman correlation test
 352 has been chosen due to the non-normality of the scores obtained through the sentiment analysis of
 353 requests, revealed by the Shapiro-Wilkinson tests. Correlograms have been also plotted to inspect the
 354 obtained distributions. Shapiro-Wilkinson test and Spearman correlation coefficients have been
 355 calculated through R (rel. 4.0) statistical language.

356 Finally, the ability of lexicons to correctly identify the severity order of each sentence has been tested
 357 comparing the order of HMA scores in respect to the order given by two of the lexicons for 150 4-tuples
 358 randomly extracted. AFINN and Jockers were chosen due to the highest correlation Spearman
 359 coefficient obtained. For each of the 4-tuples, the deviation from the correct order (detected by the
 360 HMA) has been evaluated considering the order given by the scores attributed and the order given by
 361 the three-level classification (negative, neutral, positive). For each request extracted by each 4-tuple,
 362 the correct attribution of the level given by each lexicon in respect of HMA has been evaluated. The
 363 percentage of correct attributions, partially correct attributions (shift only of a position) and wrong
 364 attributions, has been also calculated.

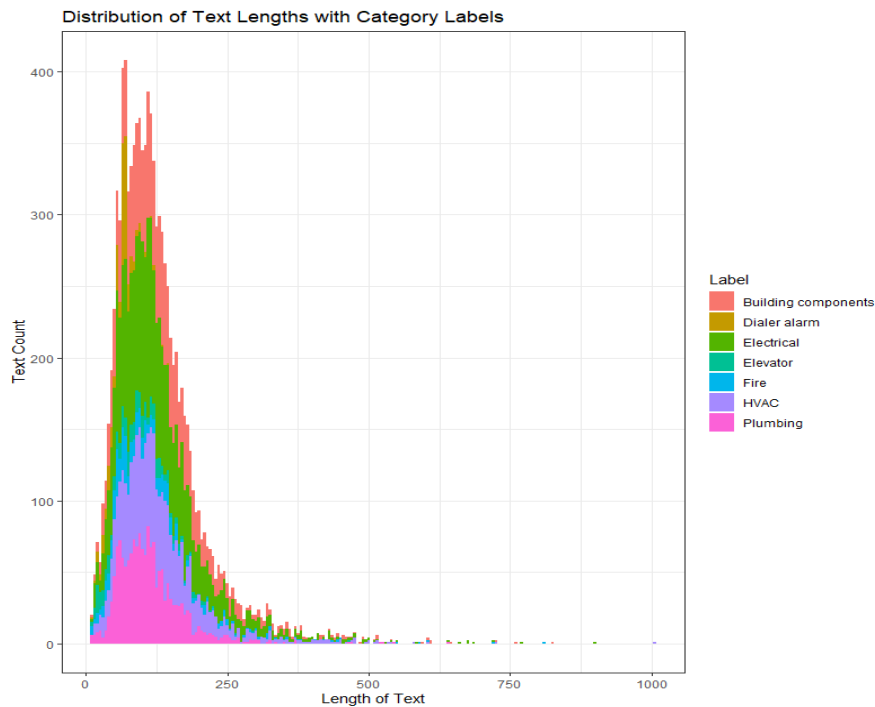
365

366 **4. Results and discussion**

367 **4.1. Term frequency analysis**

368 Each WO includes the end-user's request composed of one or more sentences, sometimes including
 369 aspects not related to the specific problem to solve. Therefore, a preliminary analysis was performed
 370 to evaluate the dimensional differences between sentences associated by technicians to specific
 371 categories. Considering the whole WOs corpus, Figure 1 shows that the end-user requests' length is
 372 not influenced by the category. Distributions are almost totally overlapped and are characterized by a
 373 typical beta left-skewed distribution. The mean length of each end-user request is 113 characters, and

374 the median is 100 (1st Quartile 70 characters; 3rd Quartile 145 characters). The “Dialer alarm” and
375 “Elevator” categories differ, being characterized by very short texts, with 66 characters and 86
376 characters as median value. It is important to underline that “Dialer alarm” is a category comprising a
377 set of e-mail messages automatically generated by the system when an alarm is detected.
378



379
380 *Figure 1 Distributions of the text lengths for each category.*
381

382 Then a TF (term frequency) analysis has been performed, to evaluate the importance of specific words
383 in the end user's maintenance requests corpus document. Words identifying buildings and parts of the
384 buildings (i.e. offices, stairs, etc...) were excluded. Figure 2 shows the TF distribution plots. The most
385 frequent words can help to identify specific categories. “Door” can help to identify building category
386 issues, “light” and “neon” (lighting) can help to identify electrical category issues, “air” can help to identify
387 HVAC category issues and “alarm” can help to identify the dialer alarm category. However, others most
388 frequent words cannot help in this task. A check of word association using Jaccard similarity (JS) of the
389 first 10 words revealed that two of them, “water” and “ceilings”, are associated with other words related
390 to different categories. E.g. “Water” is associated with “leak” (JS = 0.1686) and “bathroom” (JS =
391 0.1350), related to plumbing category, but also to “ceiling” (JS = 0.1081) and “infiltration” (JS = 0.0958),
392 close to the HVAC category.

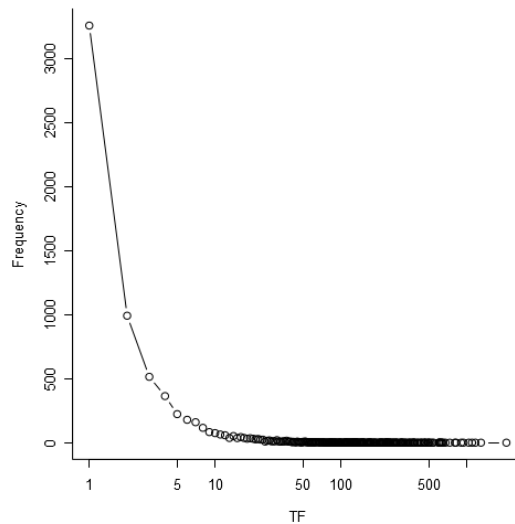


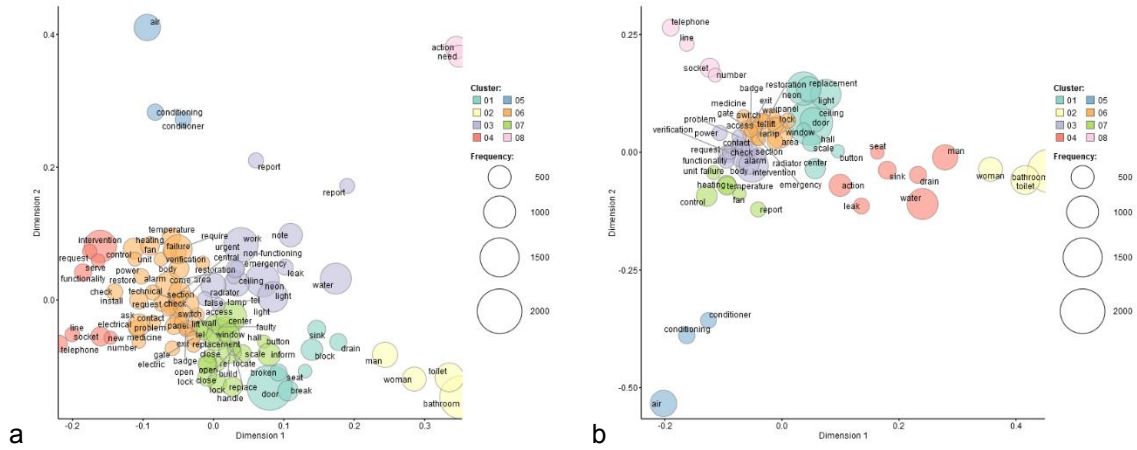
Figure 2 TF analysis of end-user requests

393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416

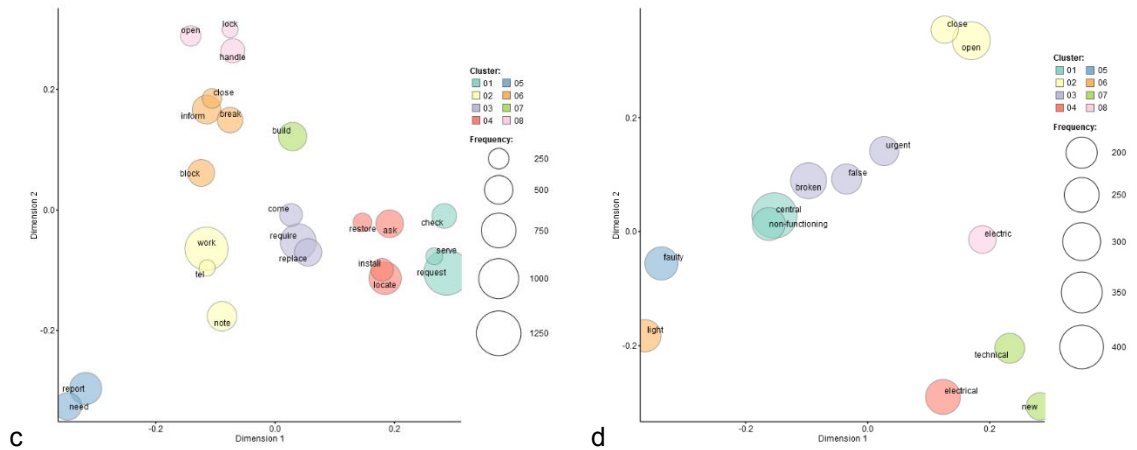
To evaluate the ability of groups of words to identify categories, a classical multidimensional scaling analysis has been performed, by filtering the corpus by nouns, verbs, adjectives and adverbs. Figure 3 shows the results of the analysis in different conditions: not filtering the bag of words (a), filtering by nouns (b) verbs (c) and adjectives/adverbs (d). Jaccard similarity has been used to represent the distance among individuals in a 2-dimension space. Bubble colour represents clusters. The bubble dimension represents the number of occurrences of each word. It's possible to observe groups of words identifying specific categories. The inclusion of all words (Figure 3a) makes it difficult to recognize clusters. However, clusters can be distinguished in an easier way by separately analyzing nouns, verbs and adjectives/adverbs, given the larger distances (Jaccard) on the plane. Figure 3b (nouns) shows that the clusters identifying maintenance WOs related to the plumbing category (cluster 6) and maintenance WOs related to HVAC systems (cluster 5) are identified thanking the analysis of the request. The cluster identifying the maintenance WOs of the "electrical category" (cluster 8) is also well identifiable. In Figure 3c (verbs), it is possible to identify the types of action required. Figure 3d (adjectives/adverbs) expresses the severity of the problem complained of.

Figure 4 represents potential relationships between groups of words. The bubble dimension represents the frequency of co-occurrence and the colours represent clusters. Through co-occurrence plots, it is possible to observe more clearly the association between words identifying categories and related clusters, and the frequency of association between words. The biggest bubbles identify the most frequent associations: "door, handle, lock", "bathroom, toilet, water, drain, sink, leak, woman, man". Co-occurrence maps also provide evidence of the association between words used to ask the intervention. The verbs "to require", "to restore", "to check" are frequently used in association with the nouns "intervention" and "functionality". "Need" and "action" words are also often used together.

417



418



419
420

Figure 3 Classical multi-dimensional scaling of words contained into end-users' maintenance requests in the whole WOs corpus: (a) all; (b) nouns; (c) verbs; (d) adjectives and adverbs. Distances are based on Jaccard similarity coefficient.

421
422
423
424

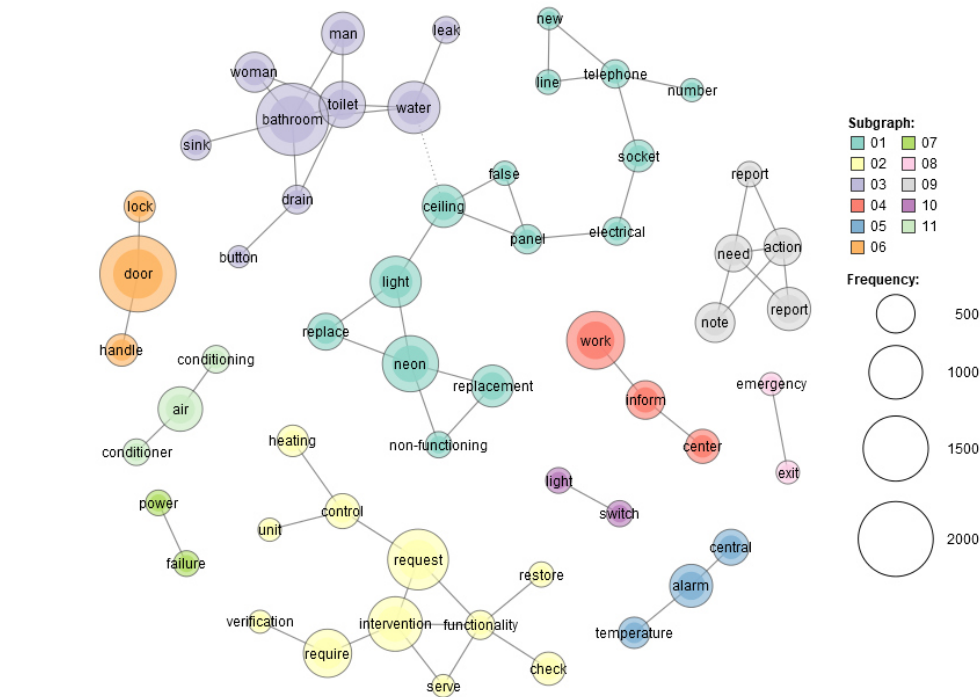
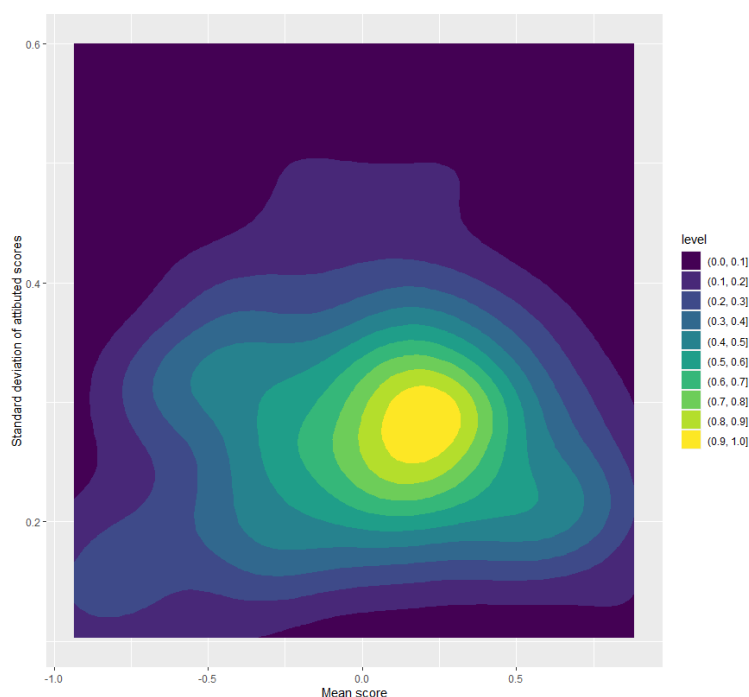


Figure 4 Co-occurrence network of terms based on their paired presence within each sentence.

4.2. HMA results and HMA-SSA comparison

425 The contingency table shows a good global agreement between the annotators and the number of
 426 sentences with score attributed by each annotator. Results strongly diverging from the mean value ($< m+2s$) is very low: 5% for the HE group, 2% for NE and 4% for LE group. Figure 4 shows this result
 427 according to a 2D kernel density of the mean and the standard deviation of the scores attributed to each
 428 sentence (-1 very negative; 1 very positive). Annotators agree almost totally on extreme (very negative
 429 or very positive) sentences. On the contrary, although the highest distribution scores can be noticed for
 430 the mean score ranging from 0.0 to 0.5, they seem to do not agree on the sentences with a mean score
 431 near the neutrality. This result is confirmed by the distribution of the mean score and the related
 432 standard deviation characterizing each sentence, as in Figure 4. in fact, standard deviations are low for
 433 sentences characterized by high positive or negative values.
 434



435
 436 *Figure 4 2D kernel density of HMA mean and standard deviation scores given by the annotators to each sentence. Colours*
 437 *represent the distribution of the scores on a scale 0-1.*
 438

439 The calculation of Krippendorff's coefficient for the thirteen annotators confirms that there is an
 440 acceptable level achieved coding the single units of analysis (sentences). In fact, $\alpha = 0.67$, thus
 441 suggesting that the final score attributed to each WO can be calculated as the mean of the scores
 442 attributed by each of the annotators. Due to the necessity to compare the gold standard (HMA) with
 443 methods characterized by numeric scores (lexicons) or level (SSA), HMA numerical scores were also
 444 converted into levels, cutting the score scale into three different levels (Negative, Neutral, Positive),
 445 characterized by the same size.

446 SSA method has been applied to extract severity level from each WO, based on a pre-defined list of
 447 high and low severity words. HMA (level scale) and SSA results have been also compared through
 448 Precision, Recall and F-score [65,66].

449 Table 4 shows that the SSA [9] method in respect to the gold standard reference (HMA) gives an F-
 450 score of 55% for Negative sentences and lower values for Neutral sentences (22%) and very low values
 451 for Positive (5%) sentences. Low SSA F-scores, especially for Neutral (medium severity) and Positive

452 (low severity) sentences, could be explained considering the high agreement reached by annotators on
453 common words expressing urgency (e.g. urgent, safety, emergency, alarm, fire), but not on words
454 expressing medium or low urgency.

455

| | Precision | Recall | F-score |
|------------------------------|------------------|---------------|----------------|
| NEG (High severity) | 0.42 | 0.81 | 55% |
| NEU (medium severity) | 0.33 | 0.17 | 22% |
| POS (low severity) | 0.29 | 0.03 | 5% |

Table 4 Precision, Recall and F-1 scores

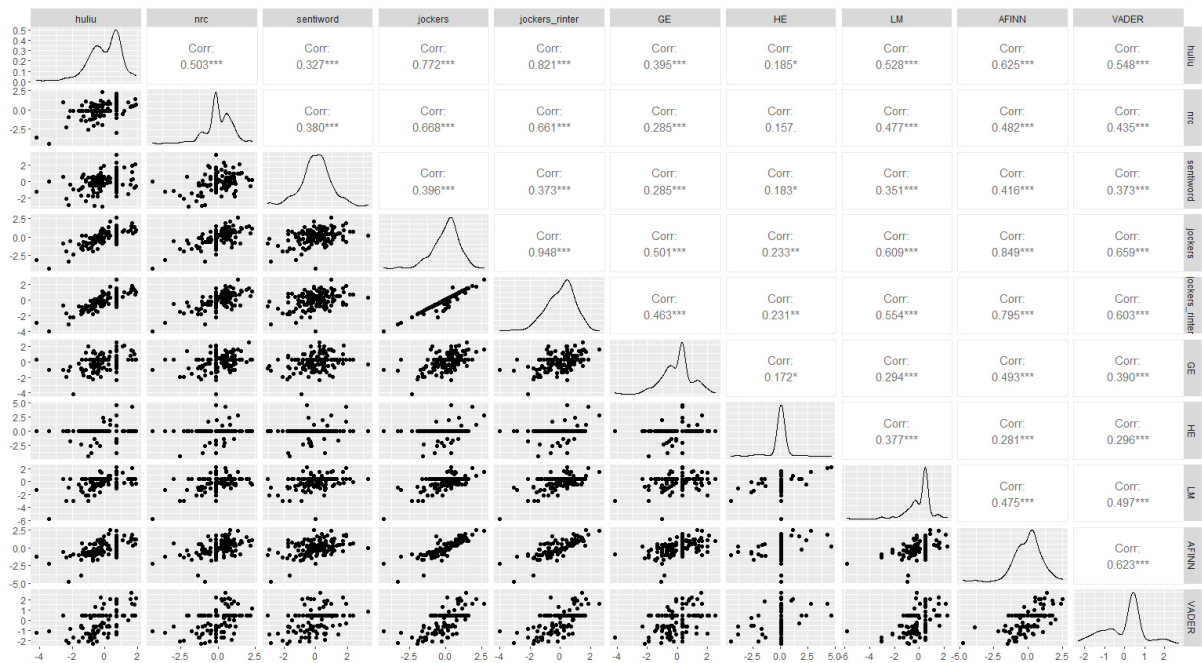
456
457

458 4.3. Lexicons comparison

459 Figure 6 shows a correlogram based on the Spearman's ρ rank correlation coefficient, where the scores
460 obtained through each lexicon are compared. At first, data were normalized to obtain scores distribution
461 characterized by mean=0 and standard deviation=1. Senticnet and QDAP lexicons were excluded due
462 to the statistical not significance of the test ($p > 0.05$).

463 As expected, the correlation coefficients are very high for those lexicons which are mainly improvements
464 of the other lexicons, i.e. in the case of AFINN, Jockers (improvement of AFINN lexicon) and Jockers-
465 Rinker (improvement of Jockers lexicon), where the spearman's ρ rank correlation coefficient is 0.949
466 (J-JR), 0.843 (J-AFINN), 0.791 (JR-AFINN). This is also the case of Jockers-Rinker (combined
467 improvement of Jockers and Hu-Liu lexicons) and Hu-Liu, where the spearman's ρ rank correlation
468 coefficient R is 0.824 (JR-HuLiu).

469 Looking at the distribution of the scores (in the diagonal of the matrix), HE and LM lexicons show a
470 consistent number of neutral requests in respect to other lexicons. This aspect is due to the intrinsic
471 characteristic of these two lexicons that contain a list of polarity annotated words for textual analysis
472 mainly in financial applications. Then, only a little number of words of these lexicons could help to
473 properly classify requests polarity. VADER lexicon also shows a significant number of WO's recognized
474 as neutral. In all these cases, the Spearman's ρ rank correlation coefficient with the other lexicons
475 remains quite low. Apart from these, the shape and the skewness of the WO's polarity score
476 distributions obtained with the other lexicons give evidence of their ability to properly represent the
477 general negative content of requests, due to the nature of the end-users' communication.



479

480 *Figure 6 Correlogram of the considered lexicons. For each pair of lexicons is reported the spearman's ρ rank correlation*
 481 *coefficient and the paired scatterplot. Senticnet and QDAP lexicons were excluded due to the statistical not significance*
 482 *of the test ($p > 0.05$).*

483

484 **4.4. HMA and Lexicon comparison**

485 Hu-Liu, NRC, Sentiword, Jockers, Jockers-Rinker, AFINN and VADER have been then compared with
 486 HMA. Senticnet, QDAP, HE and LM have been excluded considering previous results obtained
 487 analyzing the scores' distribution.

488 After preliminary tests to check the normality of the sample through the Shapiro-Wilkinson method, the
 489 Spearman correlation coefficient has been calculated.

490

| | Hu-Liu | NRC | Sentiword | Jockers | Jock_r | GE | AFINN | VADER |
|-----|--------|------|-----------|---------|--------|------|-------|-------|
| HMA | 0.21 | 0.16 | 0.25 | 0.28 | 0.25 | 0.26 | 0.28 | 0.36 |

491 *Table 4 Spearman's ρ rank correlation coefficient R of HMA in respect to the selected lexicons.*

492

493 Table 4 shows a low Spearman correlation coefficient for all the lexicons. Best results seem to be
 494 obtained by VADER, AFINN, GE and Jockers lexicons, but the correlations are weak.

495 Figure 7 shows a correlogram with a visual representation of the correlations through a scatterplot.
 496 VADER gives the highest correlation coefficients, but results are affected by many requests recognized
 497 neutral on the contrary of HMA. GE results also are affected by the same problem. AFINN (a manually
 498 annotated list of words) and Jockers (based on AFINN lexicon) give a more distributed representation
 499 even with lower correlation values.

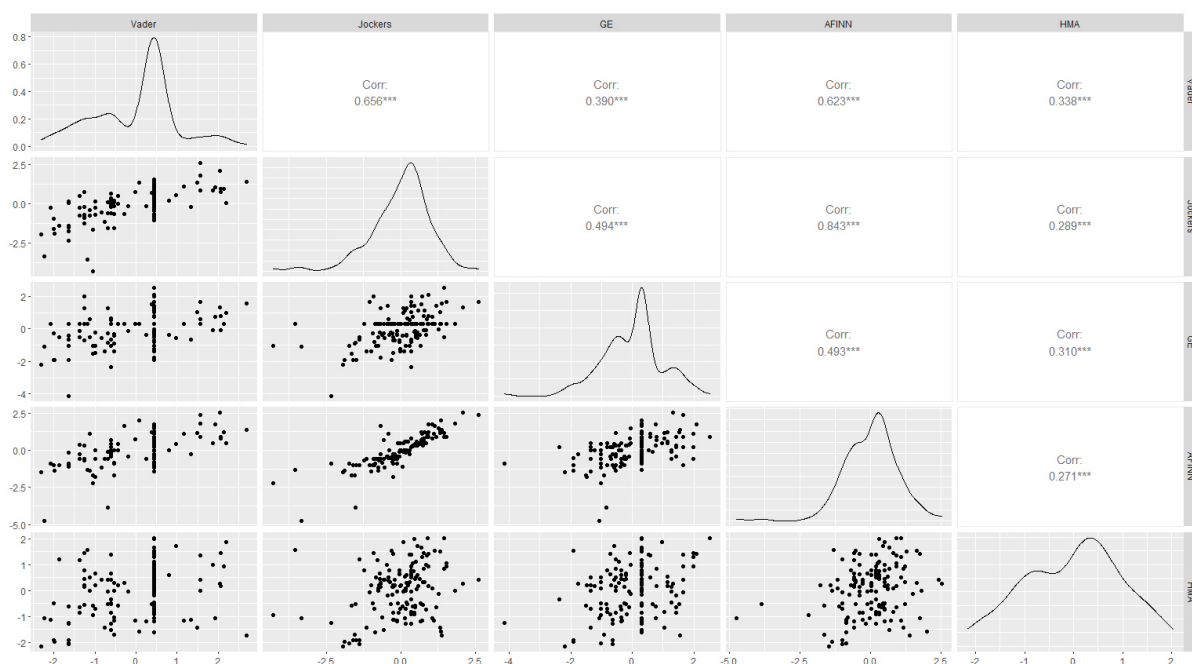


Figure 7 Correlogram showing HMA, Vader, Jockers, AFINN and GE correlations

501

502

503

504 To understand the reason for the weak correlations, a sample has been extracted and the content of
 505 each sentence was analyzed. Analysis revealed that, during HMA, annotators gave scores based on
 506 the combination of the following factors: (1) their technical knowledge of the field and their ability to
 507 properly connect “what” and “where happens”, (2) the relative importance of the component expressed
 508 by technical words, and (3) the presence of words expressing polarities (i.e. “urgent”, “alarm”,
 509 “leakage”). Indeed lexicons are able to recognize general, but not technical words, as polarized. An
 510 example is represented by the words “falling” and “ceiling”. These words express a serious problem for
 511 a technician, when they jointly occur in the request, but this connection seems to be not properly
 512 recognized by lexicons, even if they are in the same polarized cluster.

513

| Recognition | JOCKERS | | AFINN | |
|-------------|---------|-----|-------|-----|
| | Value | % | Value | % |
| Correct | 312 | 52% | 296 | 49% |
| Partial | 241 | 40% | 278 | 46% |
| Wrong | 47 | 8% | 26 | 4% |

Table 5 Score assigned by Jocker and AFINN lexicons to each sentence. Partial recognition means that a shift of 1 position has been recorded (negative instead neutral or positive instead of neutral).

514

515

516

517 Finally, further evaluations were performed to assess the incidence of the weak correlation found on
 518 the ability of lexicons to properly recognize the severity order of contemporary requests, as well as to
 519 evaluate the difference with HMA method application. These analyses were performed assuming
 520 AFINN and Jockers as the best lexicons in view of the above, basing on the three-level scale (negative,
 521 neutral, positive). According to the application of 150 4-tuples randomly extracted from the dataset,
 522 Table 5 shows the score assigned by Jocker and AFINN lexicons to each sentence and the “shift” of

523 position in respect to HMA scores. On a three-level scale, lexicons can recognize the correct severity
524 only in about 50% of the cases. These values seem to imply lower general accuracy trends in respect
525 to the results of other works on sentiment analysis approaches, in which values ranged from 60% to
526 95.5% [37]. Anyway, Table 5 also shows the moderate “shit” of position (1 position), since the result is
527 totally wrong (i.e. positive instead of negative) only in 4-8% of cases. Therefore, chosen lexicons can
528 be still used to discard the less urgent WOs, rather than selecting the most severe ones. Reasons are
529 due to the problems identified below. In particular, the analysis of the requests randomly extracted and
530 the comparison with polarity scores attributed by the lexicons confirmed that the lexicons cannot
531 correctly attribute polarity due to the influence of technical words on annotator judgement as previously
532 described.

533

534 **5. Conclusion**

535 This work shows how text mining methodologies can help to extract information and opinions from end
536 users’ maintenance requests and that, through sentiment analysis, the implicit emotion in the text of
537 each request (urgency, severity, etc...) can be powerfully mined and this information can be used to
538 take immediate or further decisions. However, the analysis of many lexicons shows that sentiment
539 analysis is a complex task, requiring a fine-tuning process to adapt lexicons to specific contexts. The
540 study shows that general lexicons cannot be applied without improvement to the field of facility
541 management. The classification by severity of end-users maintenance using a three-scale level,
542 comprising negative (high severity), neutral (mean severity), positive (low severity), gives acceptable
543 results, giving the possibility to exclude less important end-users maintenance requests. However, a
544 finer recognition is not possible without further lexicon improvements.

545 The content of each end-user’s request comprises technical words helpful to recognize the severity by
546 technicians, but not properly recognized by lexicons. This fact is confirmed by results of HMA that show
547 how these words are actually “joined” by technicians to properly recognize the severity of each end
548 user’s maintenance request. Further studies will be aimed at correlating a “combined” score based on
549 the HMA, thus moving towards the proper recognition of the polarity of technical words on “what
550 happens”, “where happens” and “which component is affected”, when joined with polarized words. In
551 this way, automatic detection of maintenance requests could be improved, and specific building use-
552 oriented methodologies could be provided to include aspects correlated to the related operational
553 features of the building itself.

554

555 **6. References**

- 556 [1] I. Errandonea, S. Beltrán, S. Arrizabalaga, Digital Twin for maintenance: A literature review,
557 *Comput. Ind.* 123 (2020). <https://doi.org/10.1016/j.compind.2020.103316>.
- 558 [2] Q. Lu, X. Xie, A. Kumar, J. Mary, Automation in Construction Digital twin-enabled anomaly
559 detection for built asset monitoring in operation and maintenance, *Autom. Constr.* 118 (2020)
560 103277. <https://doi.org/10.1016/j.autcon.2020.103277>.
- 561 [3] Y.-J. Chen, Y.-S. Lai, Y.-H. Lin, BIM-based augmented reality inspection and maintenance of
562 fire safety equipment, *Autom. Constr.* 110 (2020) 103041.
563 <https://doi.org/10.1016/j.autcon.2019.103041>.
- 564 [4] Z. Ma, Y. Ren, X. Xiang, Z. Turk, Data-driven decision-making for equipment maintenance,
565 *Autom. Constr.* 112 (2020) 103103. <https://doi.org/10.1016/j.autcon.2020.103103>.
- 566 [5] H. Yan, N. Yang, Y. Peng, Y. Ren, Data mining in the construction industry: Present status,
567 opportunities, and future trends, *Autom. Constr.* 119 (2020) 103331.
568 <https://doi.org/10.1016/j.autcon.2020.103331>.
- 569 [6] H. Burak Gunay, W. Shen, G. Newsham, Data analytics to improve building performance: A
570 critical review, *Autom. Constr.* 97 (2019) 96–109. <https://doi.org/10.1016/j.autcon.2018.10.020>.
- 571 [7] G. Bernardini, E. Di Giuseppe, Towards a user-centered and condition-based approach in
572 Building Operation and Maintenance, in: J.Littlewood, R.J.Howlett, A.Capozzoli, L.C.Jain
573 (Eds.), *Sustain. Energy Build. Proc. SEB 2019 (Series Title Smart Innov. Syst. Technol. - Vol.*
574 *163 - Ser. ISSN 2190-3018)*, 1st ed., Springer Nature Singapore Pte Ltd, 2020: pp. 327–337.
575 https://doi.org/10.1007/978-981-32-9868-2_28.
- 576 [8] G. Fernandez, I. Ahmed, “Build back better” approach to disaster recovery: Research trends
577 since 2006, *Prog. Disaster Sci.* 1 (2019) 100003. <https://doi.org/10.1016/j.pdisas.2019.100003>.
- 578 [9] R. Bortolini, N. Forcada, Analysis of building maintenance requests using a text mining
579 approach: building services evaluation, *Build. Res. Inf.* 48 (2020) 207–217.
580 <https://doi.org/10.1080/09613218.2019.1609291>.
- 581 [10] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, *A Practical Guide to Sentiment Analysis*,
582 Springer International Publishing, Cham, 2017. <https://doi.org/10.1007/978-3-319-55394-8>.
- 583 [11] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University
584 Press, 2015.
- 585 [12] M. V. Mäntylä, D. Graziotin, M. Kuutila, The evolution of sentiment analysis—A review of
586 research topics, venues, and top cited papers, *Comput. Sci. Rev.* 27 (2018) 16–32.
587 <https://doi.org/10.1016/j.cosrev.2017.10.002>.
- 588 [13] T. El-Diraby, A. Shalaby, M. Hosseini, Linking Social, Semantic and Sentiment Analyses to
589 Support Modeling Transit Customers’ Satisfaction: Towards Formal Study of Opinion
590 Dynamics, *Sustain. Cities Soc.* 49 (2019) 101578. <https://doi.org/10.1016/j.scs.2019.101578>.
- 591 [14] P. Shi, Y. Gao, Y. Shen, E. Chen, H. Chen, J. Liu, Y. Chen, Y. Xiao, K. Wang, C. Shi, B. Lu,
592 Characteristics and evaluation of the effectiveness of monitoring and control measures for the
593 first 69 Patients with COVID-19 from 18 January 2020 to 2 March in Wuxi, China, *Sustain.*
594 *Cities Soc.* 64 (2020) 102559. <https://doi.org/10.1016/j.scs.2020.102559>.

- 595 [15] Q. Zhou, Z. Xu, N.Y. Yen, User sentiment analysis based on social network information and its
596 application in consumer reconstruction intention, *Comput. Human Behav.* (2018) 0–1.
597 <https://doi.org/10.1016/j.chb.2018.07.006>.
- 598 [16] M. Marzouk, M. Enaba, Text analytics to analyze and monitor construction project contract and
599 correspondence, *Autom. Constr.* 98 (2019) 265–274.
600 <https://doi.org/10.1016/j.autcon.2018.11.018>.
- 601 [17] Z. Ding, Z. Li, C. Fan, Building energy savings: Analysis of research trends based on text
602 mining, *Autom. Constr.* 96 (2018) 398–410. <https://doi.org/10.1016/j.autcon.2018.10.008>.
- 603 [18] M.L. Loureiro, M. Alló, Sensing climate change and energy issues: Sentiment and emotion
604 analysis with social media in the U.K. and Spain, *Energy Policy.* 143 (2020).
605 <https://doi.org/10.1016/j.enpol.2020.111490>.
- 606 [19] S.M. Mohammad, P.D. Turney, Crowdsourcing a word-emotion association lexicon, *Comput.*
607 *Intell.* 29 (2013) 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>.
- 608 [20] R. Plutchik, H. Kellerman, *EMOTION - Theory, Research, and Experience Vol. 4 - The*
609 *Measurement of Emotions, I*, New York, 1989.
- 610 [21] Y. Sun, Z. Wang, B. Zhang, W. Zhao, F. Xu, J. Liu, B. Wang, Residents' sentiments towards
611 electricity price policy: Evidence from text mining in social media, *Resour. Conserv. Recycl.*
612 160 (2020) 104903. <https://doi.org/10.1016/j.resconrec.2020.104903>.
- 613 [22] X. Liu, W. Hu, Attention and sentiment of Chinese public toward green buildings based on Sina
614 Weibo, *Sustain. Cities Soc.* 44 (2019) 550–558. <https://doi.org/10.1016/j.scs.2018.10.047>.
- 615 [23] R. Bortolini, N. Forcada, Facility managers' perceptions on building performance assessment,
616 *Front. Eng. Manag.* 0 (2018) 0. <https://doi.org/10.15302/j-fem-2018010>.
- 617 [24] S. Madureira, I. Flores-Colen, J. de Brito, C. Pereira, Maintenance planning of facades in
618 current buildings, *Constr. Build. Mater.* 147 (2017) 790–802.
619 <https://doi.org/10.1016/j.conbuildmat.2017.04.195>.
- 620 [25] H.B. Gunay, W. Shen, C. Yang, Text-mining building maintenance work orders for component
621 fault frequency, *Build. Res. Inf.* 47 (2019) 518–533.
622 <https://doi.org/10.1080/09613218.2018.1459004>.
- 623 [26] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, B. Bennadji, Natural language
624 processing model for managing maintenance requests in buildings, *Buildings.* 10 (2020) 1–12.
625 <https://doi.org/10.3390/BUILDINGS10090160>.
- 626 [27] T. Sexton, M. Hodkiewicz, M.P. Brundage, T. Smoker, Benchmarking for keyword extraction
627 methodologies in maintenance work orders, *Proc. Annu. Conf. Progn. Heal. Manag. Soc.*
628 *PHM.* (2018) 1–10.
- 629 [28] R. Bardhan, M. Sunikka-Blank, A.N. Haque, Sentiment analysis as tool for gender
630 mainstreaming in slum rehabilitation housing management in Mumbai, India, *Habitat Int.* 92
631 (2019) 102040. <https://doi.org/10.1016/j.habitatint.2019.102040>.
- 632 [29] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment
633 analysis, *HLT/EMNLP 2005 - Hum. Lang. Technol. Conf. Conf. Empir. Methods Nat. Lang.*
634 *Process. Proc. Conf.* (2005) 347–354. <https://doi.org/10.3115/1220575.1220619>.

- 635 [30] S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: An enhanced lexical resource
636 for sentiment analysis and opinion mining, Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010.
637 (2010) 2200–2204.
- 638 [31] A. Esuli, F. Sebastiani, SENTIWORDNET: A high-coverage lexical resource for opinion
639 mining, Evaluation. (2007) 1–26. <http://ontotext.fbk.eu/Publications/sentiWN-TR.pdf>.
- 640 [32] S.M. Mohammad, P.D. Turney, Emotions evoked by common words and phrases: using
641 mechanical turk to create an emotion lexicon, CAAGET '10 Proc. NAACL HLT 2010 Work.
642 Comput. Approaches to Anal. Gener. Emot. Text. (2010) 26–34.
643 <http://dl.acm.org/citation.cfm?id=1860631.1860635>.
- 644 [33] S.S. Sharma, G. Dutta, SentiDraw: Using star ratings of reviews to develop domain specific
645 sentiment lexicon for polarity determination, Inf. Process. Manag. 58 (2021) 102412.
646 <https://doi.org/10.1016/j.ipm.2020.102412>.
- 647 [34] E.C. Dragut, H. Wang, P. Sistla, C. Yu, W. Meng, Polarity Consistency Checking for Domain
648 Independent Sentiment Dictionaries, IEEE Trans. Knowl. Data Eng. 27 (2015) 838–851.
649 <https://doi.org/10.1109/TKDE.2014.2339855>.
- 650 [35] E. Dragut, H. Wang, C. Yu, P. Sistla, W. Meng, Polarity Consistency Checking for Sentiment
651 Dictionaries, in: Proc. 50th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.,
652 Association for Computational Linguistics, Jeju Island, Korea, 2012: pp. 997–1005.
653 <https://www.aclweb.org/anthology/P12-1105>.
- 654 [36] T. Schmidt, M. Burghardt, An Evaluation of Lexicon-based Sentiment Analysis Techniques for
655 the Plays of Gotthold Ephraim Lessing, Proc. Second Jt. SIGHUM Work. Comput. Linguist.
656 Cult. Heritage, Soc. Sci. Humanit. Lit. (2018) 139–149.
- 657 [37] G.K. Rajput, A. Kumar, S. Kundu, A comparative study on sentiment analysis approaches and
658 methods, Proc. 2020 9th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2020. (2020) 427–
659 431. <https://doi.org/10.1109/SMART50582.2020.9337106>.
- 660 [38] J.F. Sánchez-Rada, C.A. Iglesias, Social context in sentiment analysis: Formal definition,
661 overview of current trends and framework for comparison, Inf. Fusion. 52 (2019) 344–356.
662 <https://doi.org/10.1016/j.inffus.2019.05.003>.
- 663 [39] T. Schmidt, M. Burghardt, K. Dennerlein, Sentiment annotation of historic German plays: An
664 empirical study on annotation behavior, CEUR Workshop Proc. 2155 (2018) 47–52.
- 665 [40] A. Borg, M. Boldt, Using VADER sentiment and SVM for predicting customer response
666 sentiment, Expert Syst. Appl. 162 (2020) 113746. <https://doi.org/10.1016/j.eswa.2020.113746>.
- 667 [41] C. Zhao, S. Wang, D. Li, Exploiting social and local contexts propagation for inducing Chinese
668 microblog-specific sentiment lexicons, Comput. Speech Lang. 55 (2019) 57–81.
669 <https://doi.org/10.1016/j.csl.2018.10.004>.
- 670 [42] A.C. Kim Wing, A.H. bin Mohammed, M.N. bin Abdullah, A literature review on maintenance
671 priority - conceptual framework and directions, MATEC Web Conf. 66 (2016) 00004.
672 <https://doi.org/10.1051/mateconf/20166600004>.
- 673 [43] M. Hu, B. Liu, Mining and summarizing customer reviews, KDD-2004 - Proc. Tenth ACM
674 SIGKDD Int. Conf. Knowl. Discov. Data Min. (2004) 168–177.

- 675 <https://doi.org/10.1145/1014052.1014073>.
- 676 [44] C.J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of
677 social media text, *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*. (2014) 216–225.
- 678 [45] P.J.L. Margaret Bradley, Affective norms for english words (ANEW): Instruction manual and
679 affective ratings, *Japanese J. Med. Electron. Biol. Eng.* 22 (1984) 14–15.
- 680 [46] L. Gatti, M. Guerini, M. Turchi, SentiWords: Deriving a High Precision and High Coverage
681 Lexicon for Sentiment Analysis, *IEEE Trans. Affect. Comput.* 7 (2016) 409–421.
682 <https://doi.org/10.1109/TAFFC.2015.2476456>.
- 683 [47] G.D.S. Inteligentes, D. De Ingenier, O. Araque, L. Gatti, J. Staiano, M. Guerini,
684 DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques,
685 *IEEE Trans. Affect. Comput.* (2019) 1. <https://doi.org/10.1109/TAFFC.2019.2934444>.
- 686 [48] M.S. Smith, The computer and the TAT, *J. Sch. Psychol.* 6 (1968) 206–214.
687 [https://doi.org/https://doi.org/10.1016/0022-4405\(68\)90017-4](https://doi.org/https://doi.org/10.1016/0022-4405(68)90017-4).
- 688 [49] E. Henry, Are investors influenced by how earnings press releases are written?, *J. Bus.*
689 *Commun.* 45 (2008) 363–407. <https://doi.org/10.1177/0021943608319388>.
- 690 [50] T. Loughran, B. McDonald, When Is a Liability Not a Liability? Textual Analysis, Dictionaries,
691 and 10-Ks, *J. Finance.* 41 (2011) 57–59. <https://doi.org/10.2469/dig.v41.n2.20>.
- 692 [51] R. Tyler, QDAP package: Bridging the Gap Between Qualitative Data and Quantitative
693 Analysis, (2020). <https://cran.r-project.org/package=qdap>.
- 694 [52] F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,
695 *CEUR Workshop Proc.* 718 (2011) 93–98.
- 696 [53] A. Esuli, F. Sebastiani, V.G. Moruzzi, SENTIWORDNET: A Publicly Available Lexical
697 Resource for Opinion Mining, *Proc. 5th Conf. Lang. Resour. Eval.* (2006) 417–422.
698 <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217>.
- 699 [54] E. Cambria, S. Poria, R. Bajpai, B. Schuller, SenticNet 4: A semantic resource for sentiment
700 analysis based on conceptual primitives, *COLING 2016 - 26th Int. Conf. Comput. Linguist.*
701 *Proc. COLING 2016 Tech. Pap.* (2016) 2666–2677.
- 702 [55] M. Jockers, Syuzhet: Extract sentiment and plot arcs from Text.,
703 <https://github.com/Mjockers/Syuzhet>. (2017). <https://github.com/mjockers/syuzhet>.
- 704 [56] I. Arroyo-Fernández, C.F. Méndez-Cruz, G. Sierra, J.M. Torres-Moreno, G. Sidorov,
705 Unsupervised sentence representations as word information series: Revisiting TF–IDF,
706 *Comput. Speech Lang.* 56 (2019) 107–129. <https://doi.org/10.1016/j.csl.2019.01.005>.
- 707 [57] N.H. Gabriela, R. Siantama, C.I.A. Amadea, D. Suhartono, Extractive Hotel Review
708 Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual
709 Sentiment Trends, *Procedia Comput. Sci.* 179 (2021) 558–565.
710 <https://doi.org/10.1016/j.procs.2021.01.040>.
- 711 [58] K. Higuchi, KHcoder Manual v.2 English, 2015.
- 712 [59] K. Takano, M. Ueno, J. Moriya, M. Mori, Y. Nishiguchi, F. Raes, Unraveling the linguistic
713 nature of specific autobiographical memories using a computerized classification algorithm,
714 *Behav. Res. Methods.* 49 (2017) 835–852. <https://doi.org/10.3758/s13428-016-0753-x>.

- 715 [60] S.M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for
716 20,000 English words, ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.
717 (Long Pap. 1 (2018) 174–184. <https://doi.org/10.18653/v1/p18-1017>.
- 718 [61] J.J. Louviere, T.N. Flynn, A.A.J. Marley, Best-Worst Scaling: Theory, Methods and
719 Applications, Cambridge University Press, 2015.
720 <https://books.google.it/books?id=W9uCCgAAQBAJ>.
- 721 [62] S. Kiritchenko, S.M. Mohammad, Best–Worst scaling more reliable than rating scales: A case
722 study on sentiment intensity annotation, ACL 2017 - 55th Annu. Meet. Assoc. Comput.
723 Linguist. Proc. Conf. (Long Pap. 2 (2017) 465–470. <https://doi.org/10.18653/v1/P17-2074>.
- 724 [63] M.M. Tavakoli, H. Shirouyehzad, F.H. Lotfi, S.E. Najafi, Proposing a novel heuristic algorithm
725 for university course timetabling problem with the quality of courses rendered approach; a
726 case study, Alexandria Eng. J. 59 (2020) 3355–3367.
727 <https://doi.org/10.1016/j.aej.2020.05.004>.
- 728 [64] T. Rinker, Package “sentimentr”: Calculate Text Polarity Sentiment, (2019) 59. [https://cran.r-](https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf)
729 [project.org/web/packages/sentimentr/sentimentr.pdf](https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf).
- 730 [65] P. Gonçalves, M. Araújo, F. Benevenuto, M. Cha, Comparing and combining sentiment
731 analysis methods, COSN 2013 - Proc. 2013 Conf. Online Soc. Networks. (2013) 27–37.
732 <https://doi.org/10.1145/2512938.2512951>.
- 733 [66] F.N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, F. Benevenuto, SentiBench - a
734 benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Sci. 5
735 (2016) 1–29. <https://doi.org/10.1140/epjds/s13688-016-0085-1>.
736
737

HIGHLIGHTS

- End-users' maintenance requests are studied as a source for maintenance severity ranking.
- The effectiveness of several existing Sentiment Analysis (SA) methods and a developed Human Manual Annotation (HMA) method is compared.
- About 12.000 requests for 34 months in 23 buildings of a University Campus were collected.
- HMA can better recognize the importance of technical words for maintenance severity assessment.
- Results represent a first step for future lexicon development through HMA-based methods.

Declaration of interest statement

No potential competing interest was reported by the authors.