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

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

Marco D'Orazio[1], Elisa Di Giuseppe[1], Gabriele Bernardini[1,*]

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

Abstract

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

Keywords

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

1. Introduction

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

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

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

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

More recently, natural language processing models were applied to the facility management of buildings, collecting sentiments and opinions from end-users, to improve the building operability and the cost of the management process [23,24]. Bortolini and Forcada developed a methodology, based on the TF analysis of words expressing the severity degree, to determine the typical problems that end-users complain about the building systems and their perceived severity [9]. Gunay et al. analyzed operators’ work order descriptions in Computerized Maintenance Management Systems (CMMS), extracting information about failure patterns in building systems and components [25]. The results provide insights into equipment breakdown of failure events, top system and component-level failure modes, and their occurrence frequencies. Bouabdallaoui et al. proposed a machine-learning algorithm based on Natural Language Processing (NLP) to manage day-to-day maintenance activities [26]. Sexton et al. compared NLP methodologies to extract keywords from maintenance Work orders [27]. Bardhan et al. employed two emotion lexicon databases, the Ho-Liu database [11] and the NRC emotion lexicon from semi-structured interviews and focus group discussions regarding housing management in India [28]. The author justifies the choice of two lexicons arguing that the Ho-Liu lexicon is a tool to understand the general sentiments of the documents in a binary fashion, considering only positive or negative sentiment as categorized in [11], while the NRC lexicon enables the classification of the sentiments into discreet emotions [19]. Sun et al. calculated sentiment value on energy price policies basing on polarity and intensity of sentiment words, based on the China HowNet Thesaurus [21]. The authors adopted a sentence pattern based on sentiment words, privative, degree words and rhetorical question.

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

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

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

However, despite a significant amount of research, challenging problems remain. In this context, a general and effective method for discovering and determining domain and context-dependent sentiments is still lacking [41]. It is hence necessary to preliminarily check the accuracy of proposed methodologies when applied to each specific domain to extract information about specific aspects. Then, a wide comparison between sentiment analysis techniques and HMA methods should be provided to better assess differences and similarities between them, especially when moving towards the automatic detection of the priority order in maintenance requests, which is a paramount element to support O&M [42]. Indeed, the immediate and automatic detection of the severity (importance and urgency) of any maintenance request, through text mining methodologies, could reduce the risks associated with late interventions and improve preventive maintenance strategies, providing useful information to change on-the-fly planned activities and reducing buildings' O&M costs.

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

2. Related works on sentiment analysis lexicons

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

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

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

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 or 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] .

The second group comprises lexicons useful to determine not just the binary polarity (positive versus negative), but also the strength of the sentiment expressed in text. Thus, sentiment intensity lexicons can recognize the strength of sentiment. Sentiment intensity lexicons have been further improved with disambiguation processes and mixing lexical features with rules that embody grammatical and syntactical conventions used by humans when expressing or emphasizing sentiment intensity [44]. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a sentiment intensity lexicon that combines quantitative and qualitative features. The Affective Norms for English Words (ANEW) lexicon provides a set of normative emotional ratings for 1,034 English words [45]. ANEW words have an associated sentiment valence ranging from 1-9. SentiWordNet (SWN) is an extension of WordNet in which 147,306 synsets are annotated with three numerical scores relating to positivity, negativity, and neutrality [30]. SentiWords, an high coverage lexicon for sentiment analysis based on SentiWordNet [46]. SenticNet is a publicly available semantic and affective resource for concept-level opinion and sentiment analysis [10]. The SenticNet lexicon consists of 14,244 common sense concepts such as wrath, adoration, woe, and admiration with information associated with (among other things) the concept's sentiment polarity, a numeric value on a continuous scale ranging from –1 to 1. More recently also emotion lexicons were introduced. NRC Emolex (also called NRC Word-Emotion Association Lexicon, described in [19]) classifies sentiment by mapping a large list of emotions into eight basic groups of emotions: trust (acceptance, admiration, like); fear (fear); surprise (uncertainty, amazement), sadness (sadness), disgust (dislike, hate, dis- appointment, indifference) anger (anger), anticipation (anticipation and vigilance) and joy (calmness, joy) into a four-point scale in addition to the positive and negative words [20]. Gatti et al. introduced other available emotion lexica: NRC Hashtag, NRC Affect, WordNet-Affect (wordnet extension); AffectNet; Fuzzy Affect Lexicon; Emolex; Affect; DepecheMood ++ [46]. DepecheMood++, also called DM++, is a bi-lingual lexicon (English- Italian) improvement of DepecheMood, developed in [47]. DM++ and has been compared with Hu-Liu, MPQA, NRC-Emolex, SentiWordNet lexicons in the task of emotion intensity prediction.

3. Methodology

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

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

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

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

3.2. WO's Text mining

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

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

3.3. Human manual annotation and semi-automatic human annotation

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

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

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

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

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

3.4. Sentiment and emotion analysis

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

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

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

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

3.5. Comparison methodology

HMA has been assumed as the gold standard [36] to measure the ability of the other methods to correctly evaluate the WO's severity. HMA results were expressed both on a numeric scale and on a three-level scale (negative, neutral, positive). The score conversion into a three-level scale is justified by the necessity to compare HMA with methods characterized only by a three-level scale (i.e SSA) [9]. Firstly, SSA results have been compared with the HMA according to the three ranking scales previously described. Precision, Recall and F1 measure [65,66] have been used to compare results by groups (Table 2). Recall is the ratio of the number of elements correctly classified to the number of known elements in each class. Precision is the ratio of the number of elements correctly classified to the total predicted in each class. F1 measure is the harmonic mean between both precision and recall. In detail, the precision of the negative class is computed as: $P(neg)= i/(c + f + i)$; its recall, as: $R(neg)= i/(g + h + i)$; and $F1(neg)= [2P(neg) * R(neg)] / [P(neg)+R(neg)]$.

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

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

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

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

4. Results and discussion

4.1. Term frequency analysis

Each WO includes the end-user's request composed of one or more sentences, sometimes including aspects not related to the specific problem to solve. Therefore, a preliminary analysis was performed to evaluate the dimensional differences between sentences associated by technicians to specific categories. Considering the whole WOs corpus, Figure 1 shows that the end-user requests' length is not influenced by the category. Distributions are almost totally overlapped and are characterized by a typical beta left-skewed distribution. The mean length of each end-user request is 113 characters, and the median is 100 (1st Quartile 70 characters; 3rd Quartile 145 characters). The “Dialer alarm” and “Elevator” categories differ, being characterized by very short texts, with 66 characters and 86 characters as median value. It is important to underline that “Dialer alarm” is a category comprising a set of e-mail messages automatically generated by the system when an alarm is detected.

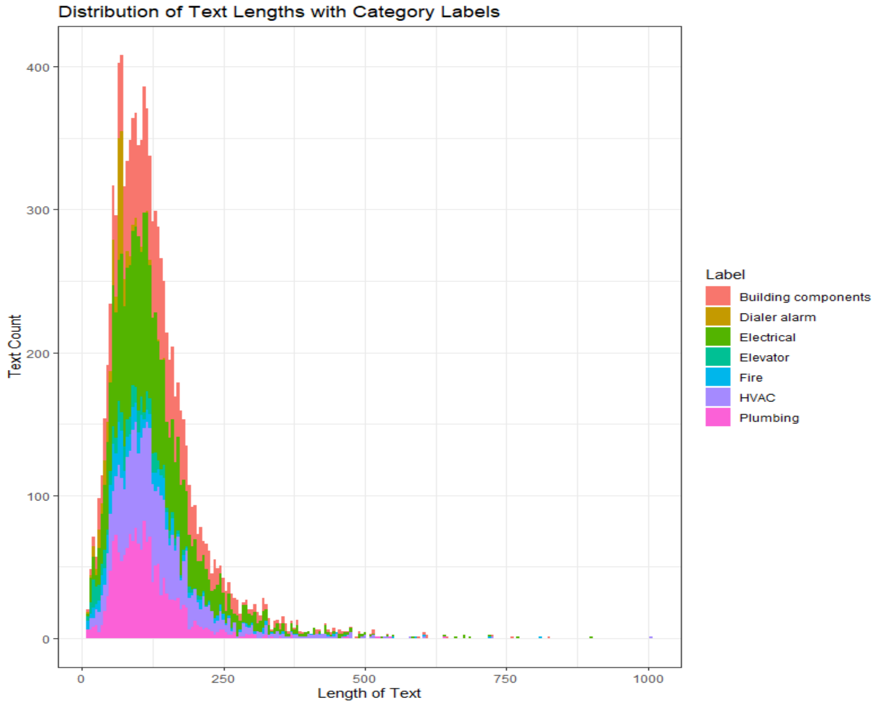

Figure 1 Distributions of the text lengths for each category.

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

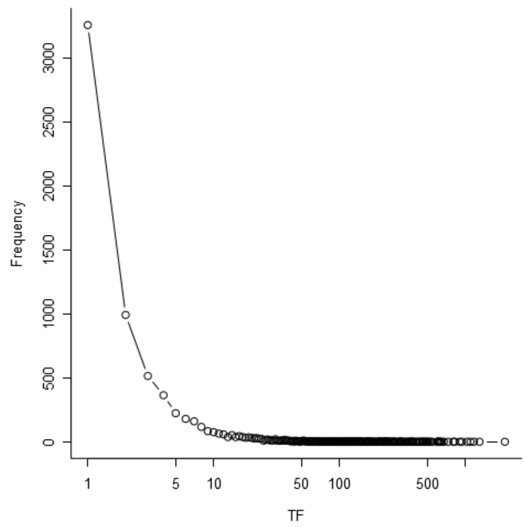

Figure 2 TF analysis of end-user requests

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.

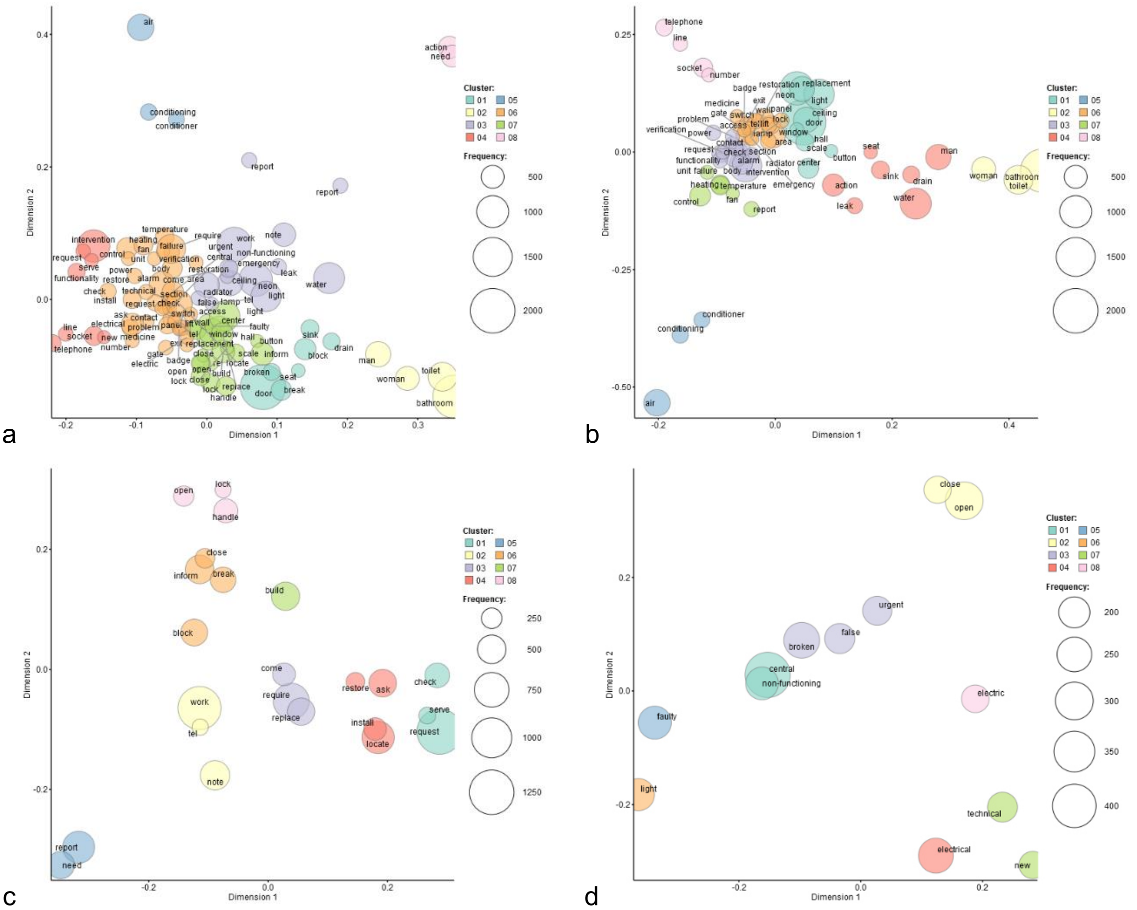

Figure 3 Classical multi-dimensional scaling of words contained into end-users' maintenance requests in the whole WOs corpus: (a) all; (b) nouns; (c) verbs; (d) adjectives and adverbs. Distances are based on Jaccard similarity coefficient.

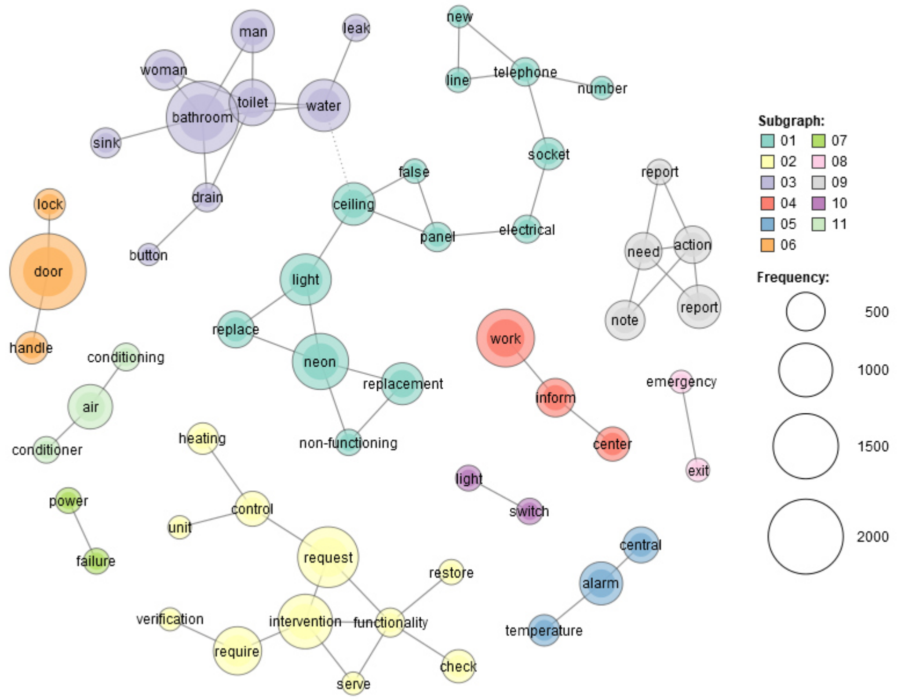

Figure 4 Co-occurrence network of terms based on their paired presence within each sentence.

4.2. HMA results and HMA-SSA comparison

The contingency table shows a good global agreement between the annotators and the number of 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 according to a 2D kernel density of the mean and the standard deviation of the scores attributed to each sentence (-1 very negative; 1 very positive). Annotators agree almost totally on extreme (very negative or very positive) sentences. On the contrary, although the highest distribution scores can be noticed for the mean score ranging from 0.0 to 0.5, they seem to do not agree on the sentences with a mean score near the neutrality. This result is confirmed by the distribution of the mean score and the related standard deviation characterizing each sentence, as in Figure 4. in fact, standard deviations are low for sentences characterized by high positive or negative values.

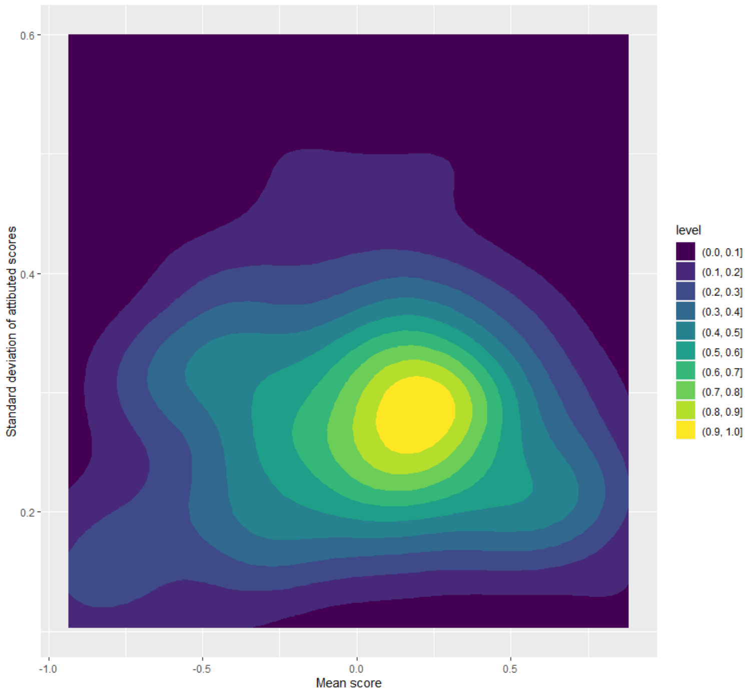

Figure 4 2D kernel density of HMA mean and standard deviation scores given by the annotators to each sentence. Colours represent the distribution of the scores on a scale 0-1.

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

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

Table 4 shows that the SSA [9] method in respect to the gold standard reference (HMA) gives an F-score of 55% for Negative sentences and lower values for Neutral sentences (22%) and very low values for Positive (5%) sentences. Low SSA F-scores, especially for Neutral (medium severity) and Positive (low severity) sentences, could be explained considering the high agreement reached by annotators on common words expressing urgency (e.g. urgent, safety, emergency, alarm, fire), but not on words expressing medium or low urgency.

	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

4.3. Lexicons comparison

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

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

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

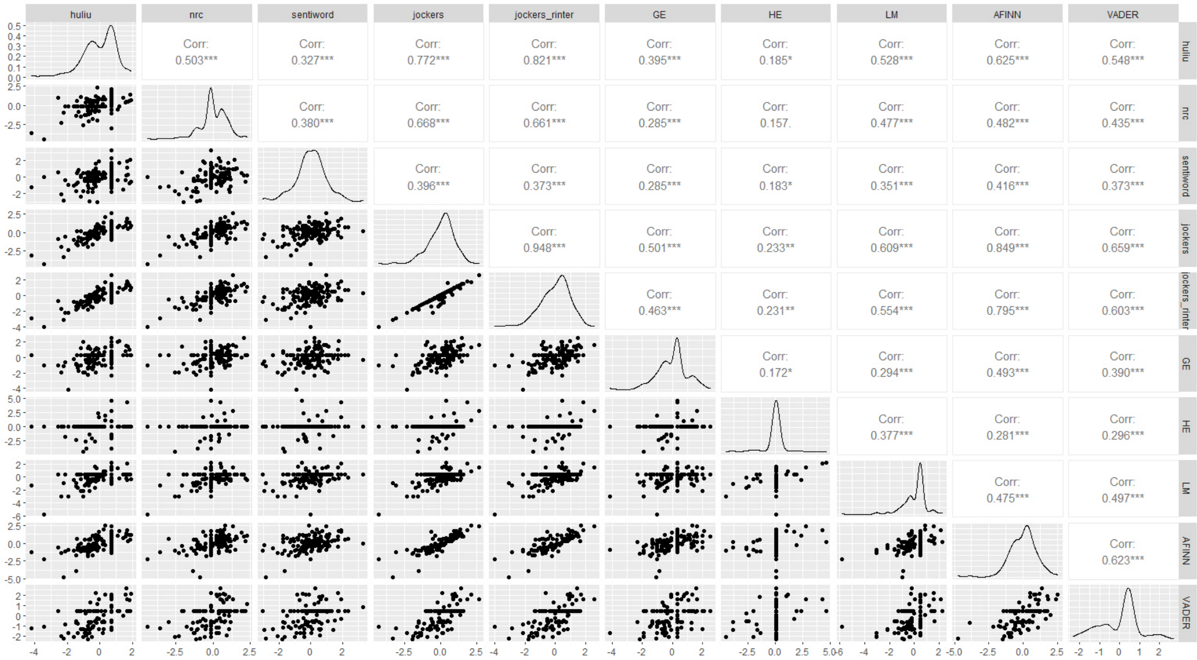

Figure 6 Correlogram of the considered lexicons. For each pair of lexicons is reported the spearman's ρ rank correlation coefficient and the paired scatterplot. Senticnet and QDAP lexicons were excluded due to the statistical not significance of the test ($p > 0.05$).

4.4. HMA and Lexicon comparison

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

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

	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

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

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

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

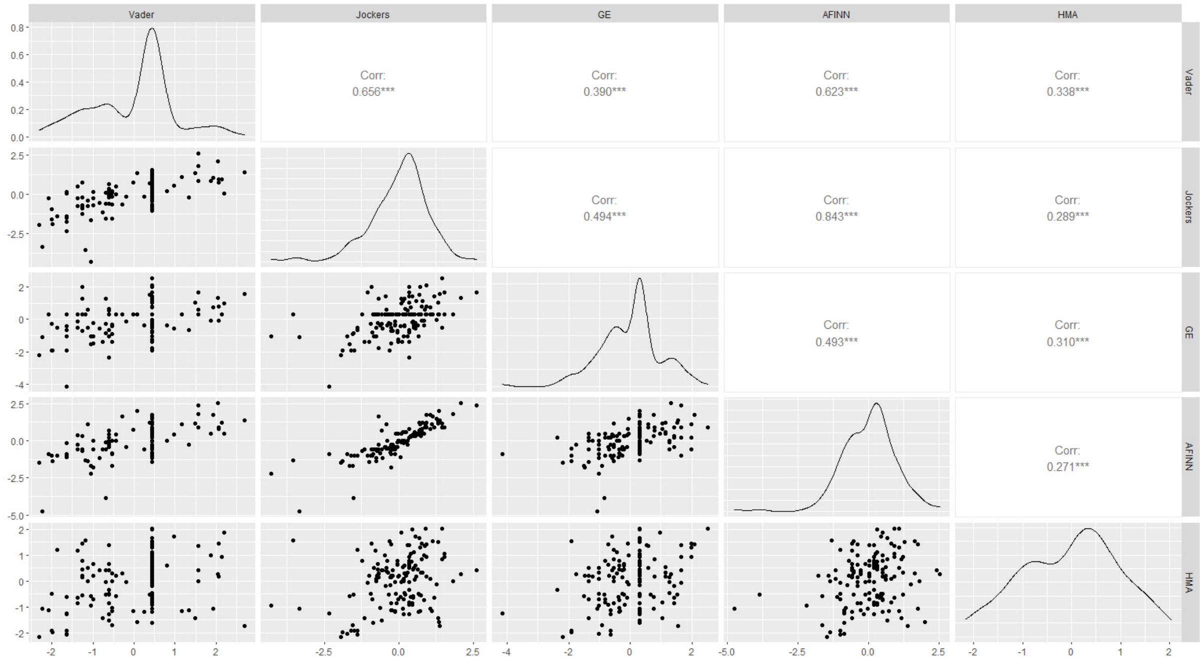

Figure 7 Correlogram showing HMA, Vader, Jockers, AFINN and GE correlations

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

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

*Table 5 S*core 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).

Finally, further evaluations were performed to assess the incidence of the weak correlation found on the ability of lexicons to properly recognize the severity order of contemporary requests, as well as to evaluate the difference with HMA method application. These analyses were performed assuming AFINN and Jockers as the best lexicons in view of the above, basing on the three-level scale (negative, neutral, positive). According to the application of 150 4-tuples randomly extracted from the dataset, Table 5 shows the score assigned by Jocker and AFINN lexicons to each sentence and the "shift" of position in respect to HMA scores. On a three-level scale, lexicons can recognize the correct severity only in about 50% of the cases. These values seem to imply lower general accuracy trends in respect to the results of other works on sentiment analysis approaches, in which values ranged from 60% to 95.5% [37]. Anyway, Table 5 also shows the moderate “shit” of position (1 position), since the result is totally wrong (i.e. positive instead of negative) only in 4-8% of cases. Therefore, chosen lexicons can be still used to discard the less urgent WOs, rather than selecting the most severe ones. Reasons are due to the problems identified below. In particular, the analysis of the requests randomly extracted and the comparison with polarity scores attributed by the lexicons confirmed that the lexicons cannot correctly attribute polarity due to the influence of technical words on annotator judgement as previously described.

5. Conclusion

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

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

6. References

[1] I. Errandonea, S. Beltrán, S. Arrizabalaga, Digital Twin for maintenance: A literature review, Comput. Ind. 123 (2020). https://doi.org/10.1016/j.compind.2020.103316.

[2] Q. Lu, X. Xie, A. Kumar, J. Mary, Automation in Construction Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance, Autom. Constr. 118 (2020) 103277. https://doi.org/10.1016/j.autcon.2020.103277.

[3] Y.-J. Chen, Y.-S. Lai, Y.-H. Lin, BIM-based augmented reality inspection and maintenance of fire safety equipment, Autom. Constr. 110 (2020) 103041. https://doi.org/10.1016/j.autcon.2019.103041.

[4] Z. Ma, Y. Ren, X. Xiang, Z. Turk, Data-driven decision-making for equipment maintenance, Autom. Constr. 112 (2020) 103103. https://doi.org/10.1016/j.autcon.2020.103103.

[5] H. Yan, N. Yang, Y. Peng, Y. Ren, Data mining in the construction industry: Present status, opportunities, and future trends, Autom. Constr. 119 (2020) 103331. https://doi.org/10.1016/j.autcon.2020.103331.

[6] H. Burak Gunay, W. Shen, G. Newsham, Data analytics to improve building performance: A critical review, Autom. Constr. 97 (2019) 96–109. https://doi.org/10.1016/j.autcon.2018.10.020.

[7] G. Bernardini, E. Di Giuseppe, Towards a user-centered and condition-based approach in Building Operation and Maintenance, in: J.Littlewood, R.J.Howlett, A.Capozzoli, L.C.Jain (Eds.), Sustain. Energy Build. Proc. SEB 2019 (Series Title Smart Innov. Syst. Technol. - Vol. 163 - Ser. ISSN 2190-3018), 1st ed., Springer Nature Singapore Pte Ltd, 2020: pp. 327–337. https://doi.org/10.1007/978-981-32-9868-2_28.

[8] G. Fernandez, I. Ahmed, “Build back better” approach to disaster recovery: Research trends since 2006, Prog. Disaster Sci. 1 (2019) 100003. https://doi.org/10.1016/j.pdisas.2019.100003.

[9] R. Bortolini, N. Forcada, Analysis of building maintenance requests using a text mining approach: building services evaluation, Build. Res. Inf. 48 (2020) 207–217. https://doi.org/10.1080/09613218.2019.1609291.

[10] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, A Practical Guide to Sentiment Analysis, Springer International Publishing, Cham, 2017. https://doi.org/10.1007/978-3-319-55394-8.

[11] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.

[12] M. V. Mäntylä, D. Graziotin, M. Kuutila, The evolution of sentiment analysis—A review of research topics, venues, and top cited papers, Comput. Sci. Rev. 27 (2018) 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002.

[13] T. El-Diraby, A. Shalaby, M. Hosseini, Linking Social, Semantic and Sentiment Analyses to Support Modeling Transit Customers’ Satisfaction: Towards Formal Study of Opinion Dynamics, Sustain. Cities Soc. 49 (2019) 101578. https://doi.org/10.1016/j.scs.2019.101578.

[14] P. Shi, Y. Gao, Y. Shen, E. Chen, H. Chen, J. Liu, Y. Chen, Y. Xiao, K. Wang, C. Shi, B. Lu, Characteristics and evaluation of the effectiveness of monitoring and control measures for the first 69 Patients with COVID-19 from 18 January 2020 to 2 March in Wuxi, China, Sustain. Cities Soc. 64 (2020) 102559. https://doi.org/10.1016/j.scs.2020.102559.

[15] Q. Zhou, Z. Xu, N.Y. Yen, User sentiment analysis based on social network information and its application in consumer reconstruction intention, Comput. Human Behav. (2018) 0–1. https://doi.org/10.1016/j.chb.2018.07.006.

[16] M. Marzouk, M. Enaba, Text analytics to analyze and monitor construction project contract and correspondence, Autom. Constr. 98 (2019) 265–274. https://doi.org/10.1016/j.autcon.2018.11.018.

[17] Z. Ding, Z. Li, C. Fan, Building energy savings: Analysis of research trends based on text mining, Autom. Constr. 96 (2018) 398–410. https://doi.org/10.1016/j.autcon.2018.10.008.

[18] M.L. Loureiro, M. Alló, Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain, Energy Policy. 143 (2020). https://doi.org/10.1016/j.enpol.2020.111490.

[19] S.M. Mohammad, P.D. Turney, Crowdsourcing a word-emotion association lexicon, Comput. Intell. 29 (2013) 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[20] R. Plutchik, H. Kellerman, EMOTION - Theory, Research, and Experience Vol. 4 - The Measurement of Emotions, I, New York, 1989.

[21] Y. Sun, Z. Wang, B. Zhang, W. Zhao, F. Xu, J. Liu, B. Wang, Residents' sentiments towards electricity price policy: Evidence from text mining in social media, Resour. Conserv. Recycl. 160 (2020) 104903. https://doi.org/10.1016/j.resconrec.2020.104903.

[22] X. Liu, W. Hu, Attention and sentiment of Chinese public toward green buildings based on Sina Weibo, Sustain. Cities Soc. 44 (2019) 550–558. https://doi.org/10.1016/j.scs.2018.10.047.

[23] R. Bortolini, N. Forcada, Facility managers' perceptions on building performance assessment, Front. Eng. Manag. 0 (2018) 0. https://doi.org/10.15302/j-fem-2018010.

[24] S. Madureira, I. Flores-Colen, J. de Brito, C. Pereira, Maintenance planning of facades in current buildings, Constr. Build. Mater. 147 (2017) 790–802. https://doi.org/10.1016/j.conbuildmat.2017.04.195.

[25] H.B. Gunay, W. Shen, C. Yang, Text-mining building maintenance work orders for component fault frequency, Build. Res. Inf. 47 (2019) 518–533. https://doi.org/10.1080/09613218.2018.1459004.

[26] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, B. Bennadji, Natural language processing model for managing maintenance requests in buildings, Buildings. 10 (2020) 1–12. https://doi.org/10.3390/BUILDINGS10090160.

[27] T. Sexton, M. Hodkiewicz, M.P. Brundage, T. Smoker, Benchmarking for keyword extraction methodologies in maintenance work orders, Proc. Annu. Conf. Progn. Heal. Manag. Soc. PHM. (2018) 1–10.

[28] R. Bardhan, M. Sunikka-Blank, A.N. Haque, Sentiment analysis as tool for gender mainstreaming in slum rehabilitation housing management in Mumbai, India, Habitat Int. 92 (2019) 102040. https://doi.org/10.1016/j.habitatint.2019.102040.

[29] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, HLT/EMNLP 2005 - Hum. Lang. Technol. Conf. Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. (2005) 347–354. https://doi.org/10.3115/1220575.1220619.

[30] S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010. (2010) 2200–2204.

[31] A. Esuli, F. Sebastiani, SENTIWORDNET: A high-coverage lexical resource for opinion mining, Evaluation. (2007) 1–26. http://ontotext.fbk.eu/Publications/sentiWN-TR.pdf.

[32] S.M. Mohammad, P.D. Turney, Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon, CAAGET ’10 Proc. NAACL HLT 2010 Work. Comput. Approaches to Anal. Gener. Emot. Text. (2010) 26–34. http://dl.acm.org/citation.cfm?id=1860631.1860635.

[33] S.S. Sharma, G. Dutta, SentiDraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination, Inf. Process. Manag. 58 (2021) 102412. https://doi.org/10.1016/j.ipm.2020.102412.

[34] E.C. Dragut, H. Wang, P. Sistla, C. Yu, W. Meng, Polarity Consistency Checking for Domain Independent Sentiment Dictionaries, IEEE Trans. Knowl. Data Eng. 27 (2015) 838–851. https://doi.org/10.1109/TKDE.2014.2339855.

[35] E. Dragut, H. Wang, C. Yu, P. Sistla, W. Meng, Polarity Consistency Checking for Sentiment Dictionaries, in: Proc. 50th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap., Association for Computational Linguistics, Jeju Island, Korea, 2012: pp. 997–1005. https://www.aclweb.org/anthology/P12-1105.

[36] T. Schmidt, M. Burghardt, An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing, Proc. Second Jt. SIGHUM Work. Comput. Linguist. Cult. Heritage, Soc. Sci. Humanit. Lit. (2018) 139–149.

[37] G.K. Rajput, A. Kumar, S. Kundu, A comparative study on sentiment analysis approaches and methods, Proc. 2020 9th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2020. (2020) 427–431. https://doi.org/10.1109/SMART50582.2020.9337106.

[38] J.F. Sánchez-Rada, C.A. Iglesias, Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison, Inf. Fusion. 52 (2019) 344–356. https://doi.org/10.1016/j.inffus.2019.05.003.

[39] T. Schmidt, M. Burghardt, K. Dennerlein, Sentiment annotation of historic German plays: An empirical study on annotation behavior, CEUR Workshop Proc. 2155 (2018) 47–52.

[40] A. Borg, M. Boldt, Using VADER sentiment and SVM for predicting customer response sentiment, Expert Syst. Appl. 162 (2020) 113746. https://doi.org/10.1016/j.eswa.2020.113746.

[41] C. Zhao, S. Wang, D. Li, Exploiting social and local contexts propagation for inducing Chinese microblog-specific sentiment lexicons, Comput. Speech Lang. 55 (2019) 57–81. https://doi.org/10.1016/j.csl.2018.10.004.

[42] A.C. Kim Wing, A.H. bin Mohammed, M.N. bin Abdullah, A literature review on maintenance priority - conceptual framework and directions, MATEC Web Conf. 66 (2016) 00004. https://doi.org/10.1051/matecconf/20166600004.

[43] M. Hu, B. Liu, Mining and summarizing customer reviews, KDD-2004 - Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2004) 168–177.

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

[60] S.M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words, ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap. 1 (2018) 174–184. https://doi.org/10.18653/v1/p18-1017.

[61] J.J. Louviere, T.N. Flynn, A.A.J. Marley, Best-Worst Scaling: Theory, Methods and Applications, Cambridge University Press, 2015. https://books.google.it/books?id=W9uCCgAAQBAJ.

[62] S. Kiritchenko, S.M. Mohammad, Best–Worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap. 2 (2017) 465–470. https://doi.org/10.18653/v1/P17-2074.

[63] M.M. Tavakoli, H. Shirouyehzad, F.H. Lotfi, S.E. Najafi, Proposing a novel heuristic algorithm for university course timetabling problem with the quality of courses rendered approach; a case study, Alexandria Eng. J. 59 (2020) 3355–3367. https://doi.org/10.1016/j.aej.2020.05.004.

[64] T. Rinker, Package “sentimentr”: Calculate Text Polarity Sentiment, (2019) 59. https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf.

[65] P. Gonçalves, M. Araújo, F. Benevenuto, M. Cha, Comparing and combining sentiment analysis methods, COSN 2013 - Proc. 2013 Conf. Online Soc. Networks. (2013) 27–37. https://doi.org/10.1145/2512938.2512951.

[66] F.N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, F. Benevenuto, SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Sci. 5 (2016) 1–29. https://doi.org/10.1140/epjds/s13688-016-0085-1.

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.