Knowledge and Information Systems (2020) 62:4445–4480
https://doi.org/10.1007/s10115-020-01497-6

SURVEY PAPER

Sentiment lexicons and non-English languages: a survey

Mohammed Kaity[1] · **Vimala Balakrishnan**[1]

Received: 10 October 2018 / Accepted: 14 July 2020 / Published online: 22 July 2020

Abstract
The ever-increasing number of Internet users and online services, such as Amazon, Twitter and Facebook has rapidly motivated people to not just transact using the Internet but to also voice their opinions about products, services, policies, etc. Sentiment analysis is a field of study to extract and analyze public views and opinions. However, current research within this field mainly focuses on building systems and resources using the English language. The primary objective of this study is to examine existing research in building sentiment lexicon systems and to classify the methods with respect to non-English datasets. Additionally, the study also reviewed the tools used to build sentiment lexicons for non-English languages, ranging from those using machine translation to graph-based methods. Shortcomings are highlighted with the approaches along with recommendations to improve the performance of each approach and areas for further study and research.

Keywords Sentiment analysis · Sentiment Lexicon · Lexicon-based · Multilingual sentiment analysis

1 Introduction

The Internet has enabled users to expose views and opinions regarding products, social issues, policies, and much more. Thus, the Internet has rapidly evolved into a massive data warehouse consisting of user opinions and emotions [1, 2]. Sentiment analysis is a field of study that refers to analyzing, interpreting and evaluating opinions. It is considered to be one of the most popular research areas using NLP techniques, text analysis and computational linguistics to identify text polarity, either as positive, negative or neutral [3]. Due to the urgent need to understand user trends on a particular subject, sentiment analysis has fast become one of the most critical and value-added research areas over the past few years [1, 4]. Sentiment analysis systems are necessary tools that help to analyze and interpret enormous amounts of data and information thereby identifying and extracting user's opinions

✉ Vimala Balakrishnan
vimala.balakrishnan@um.edu.my

Mohammed Kaity
moh.kaity@gmail.com

[1] Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

and emotions [1, 5]. Despite the enormity of texts available for multiple languages, the focus of most sentiment analysis studies has been primarily on the English language [5]. Sentiment analysis continues to be a major research domain to explore, acknowledging the many challenges and language difficulties that exist but does, however, demonstrate great promise [5]. Sentiment classification consists of two broad categories: lexicon-based and machine learning-based classifications [6]. The classifiers based on sentiment lexicons (i.e., a list of semantic polar words) are lexicon-based or rule-based classifiers, while machine learning-based classifiers depend on training datasets or annotated corpus. Lexicon-based sentiment analysis for any language is dependent on sentiment lexicons that are produced manually or semiautomatically [5] and commonly stored as dictionaries, thesaurus or corpora [7]. Although generating sentiment lexicons manually is very time-consuming, it is deemed to be more accurate than other methods due to the involvement of human experts [8–10]. Conversely, the semiautomatic methods combine manual and automated approaches to extract the sentiment lexicon words [6]. Creating the necessary sentiment analysis resources and making these resources available enables the construction of training data for sentiment classification tasks or the creation of rule-based sentiment analysis. The creation of sentiment lexicons, dictionaries and corpus is called resource building [11, 12], which is crucial to build any sentiment analysis system [13]. Resource building could help to enhance sentiment analysis activities; though in reality, this is not a sentiment analysis activity [11].

1.1 Motivation

Data are recognized as an asset and influential resource, enabling businesses to analyze views and opinions of users both in a positive and negative manner. Currently, most available resources are predominantly in English [14–17], hence making it difficult to perform and explore sentiment analysis in other languages accurately. Generating sentiment lexicons in other languages manually is also very time-consuming and expensive [13]. In the last decade, several studies have been carried out to examine the lack of multilingual subjectivity resources [5], and thus, resulting in the development of several lexicons in other languages, including Arabic [18], Chinese [19], Korean [20], French [21], Romanian [7], Swedish [22], Polish [23], etc. Therefore, the motivation of this paper is to study and classify the approaches, methods and tools employed to build sentiment lexicons for non-English languages.

1.2 Related work

Most of the published reviews in sentiment lexicons do not address the actual construction process of sentiment lexicons. Furthermore, the literature overlooks the need to discuss or make mention of the importance of multilingual sentiment analysis regarding the scarcity of sentiment resources for non-English languages. Ravi and Ravi [24] conducted a review analysis on sentiment analysis papers published from 2002 to 2014, focusing on sentiment classification approaches, tasks carried out and applications of algorithms. While their reviews comprised of papers published using English sentiment lexicons, our paper takes a closer look at studies in other languages.

Authors such as Lo et al. [5] highlighted some of the challenges in multilingual sentiment analysis and described techniques and tools that can be adopted for non-English languages. However, the studies examined did not address the construction of sentiment analysis resources such as lexicons and corpus. On the other hand, Medhat et al. [11] provided an extensive overview of existing algorithms and applications in sentiment analysis;

however, they did not focus on the techniques used in building lexicons. Therefore, to fill the missing gaps identified above, this study aims to provide an accurate classification of the current approaches based on their data sources focusing on sentiment lexicons for non-English languages.

1.3 Objectives

The primary objectives of this study are to define the approaches and methods used to construct non-English sentiment lexicons and to provide an overview of the resources required to build the lexicons. In this paper, a unique perspective of sentiment lexicon approaches is presented based on the type of resource required. This study aims:

1. To identify the data resources to build non-English sentiment lexicon;
2. To identify details of the methods used for building polarity lexicons for non-English languages.

This paper is organised as follows: Sect. 2 introduces sentiment lexicons, followed by a brief explanation on studies that have used English sentiment lexicons as resources for other languages. Section 3 presents the methods employed to build sentiment lexicons for non-English languages, applying three basic approaches: dictionary-based approaches; corpus-based approaches; and human-based computing approaches. Section 4 provides challenges and open issues, and finally, the paper is concluded in Sect. 5.

2 Sentiment lexicons

A sentiment lexicon is one of the most valuable resources of sentiment analysis for any language [10, 22, 25]. They are vital resources for both lexicon-based and machine-based learning approaches [13], with many researchers leveraging sentiment lexicons to produce unsupervised sentiment models or as training features to train machine learning algorithms in supervised approaches [26]. A sentiment lexicon is a collection of words (also known as polar or opinion words) associated with their sentiment orientation, that is, positive or negative [10, 11]. Examples of positive sentiment words are *wonderful, beautiful*, and *amazing*. In contrast, examples of negative sentiment words are *awful, poor* and *bad*.

Unfortunately, very few sentiment lexicons are available and accessible on the Web [5]. Some sentiment lexicons consist of a single file containing a list of words and their associations with sentiments (negative and positive), divided into two columns where the first column contains the words (or terms) and the second column contains the polarity in the form of (positive, negative), (0,1) or (1,− 1) [27], and some included the word strength as well. On the other hand, other researchers divided sentiment lexicons into two individual files, with one containing positive words and the other containing negative words (e.g., Bing Liu's sentiment lexicon [16]). The sentiment orientation (polarity value) may be represented in various forms, some of which are:

- A real value indicating sentiment strength in a range such as (−1 – + 1),
- Fixed categorization into positive or negative,
- A restricted number of ranking sets such as strongly positive, positive, neutral, negative, strongly negative [10].

Some sentiment lexicons also provide the part-of-speech (POS) for each word, whereas others provide information about the strength of the polarity [17]. Table 1 illustrates a frag-

Table 1 A fragment of the MPQA subjectivity lexicon

No.	Strength	Length	Word	POS	Stemmed	Polarity
1	type = weaksubj	len = 1	word1 = abandoned	pos1 = adj	stemmed1 = n	priorpolarity = negative
…						
3145	type = strongsubj	len = 1	word1 = gaily	pos1 = adverb	stemmed1 = n	priorpolarity = positive
3146	type = weaksubj	len = 1	word1 = gain	pos1 = noun	stemmed1 = n	priorpolarity = positive
3147	type = weaksubj	len = 1	word1 = gain	pos1 = verb	stemmed1 = y	priorpolarity = positive
3148	type = strongsubj	len = 1	word1 = gainful	pos1 = adj	stemmed1 = n	priorpolarity = positive
3149	type = strongsubj	len = 1	word1 = gainfully	pos1 = adverb	stemmed1 = n	priorpolarity = positive
…						
8221	type = strongsubj	len = 1	word1 = zest	pos1 = noun	stemmed1 = n	priorpolarity = positive

http://sentiment.christopherpotts.net/lexicons.html

ment of the Multi-Perspective Question Answering subjectivity lexicon (MPQA) [17], along with the details on the word strength, length, POS, stemmed and polarity values for each MPQA entry.

2.1 English sentiment lexicons as resources for other languages

Numerous researchers rely on sentiment lexicons available in English that have been built manually and more accurately [28]. These English sentiment lexicons have greatly helped in saving time and effort in building new sentiment lexicons for non-English languages [21]. Popular English sentiment lexicons such as SentiWordNet [29, 30], SenticNet [14], and Opinion Lexicon [16] have been used in many approaches to build non-English sentiment lexicons to improve the performance of sentiment classification. For example, SentiWordNet is a publicly available lexical resource for sentiment analysis. It is built by associating each WordNet synset to one of three categories: positive (Pos), negative (Neg) and neutral (Obj). SentiWordNet indicates the degree of each term using numerical scores ranging from 0.0 to 1.0 [4, 28, 29]. Nevertheless, like other lexicons, SentiWordNet contains some noise considering not all polarity values assigned to the terms are accurate. Moreover, some terms do not have a polarity value whereas some have conflicting values [4, 14, 31]. For example, the term "cruelly" has two polarity entries in SentiWordNet, that is, Pos = 0.125; Neg = 0, and Pos = 0; Neg = 0.125 [4, 14]. In addition, SentiWordNet also assigns polarity at the syntactic level, whereby the polarities of the words are ranked according to their parts of speech, that is, noun, adjective, verb and adverb. These are represented as "*n*", "*a*," "*v*" and "*r*," respectively [32].

Likewise, SenticNet [14] is a publicly available resource that was built by exploiting artificial intelligent and semantic Web techniques and based on a new dimensionality-reduction approach to infer the polarity of common sense concepts. Cambria et al. [14] used SenticNet to determine sentiments from text extracted from the Internet. The authors utilized

the Hourglass model [33], in which the sentiments were organized around four independent dimensions: pleasantness, attention, sensitivity and aptitude that make up the total emotional state of the mind [14, 31]. They defined concept polarity as an algebraic sum of the Hourglass categorization model's sentic labels, where Concept Polarity = ((Pleasantness + |Attention| − |Sensitivity| + Aptitude)/9) [14]. Table 2 provides a short overview of some English lexicons that are widely used to produce non-English sentiment lexicons [34].

3 Methods to build sentiment lexicons for non-English languages

In this section, the classifications and approach to build non-English sentiment lexicons are examined. This is supported by several studies illustrating each approach, and the identification of applied languages. Although the performance of sentiment analysis systems mainly depends on the coverage and the accuracy of the sentiment lexicon used, many languages have not received adequate attention for building lexicons [39]. Thus, the current sentiment lexicons available to the public have not achieved the acceptable level of precision required.

Methods for building sentiment lexicons vary from being completely manual, semiautomatic, to limited automatic approaches [22]. In this study, strategies are divided and used to construct sentiment lexicons according to the type of source used. Accordingly, there are three sources employed to build sentiment lexicons: preexisting lexicons, target language corpus and target language native speakers (i.e., humans) as shown in Table 3. Figure 1 graphically illustrates the approaches used to build sentiment lexicons for non-English languages.

The following presents a number of works carried out for each approach with the identification of the languages used. Comparisons were made between the works done within the span of 2006 and 2018 focusing on building non-English sentiment lexicons. The studies were compared in terms of approach, method, languages, data sources, technique, domain, the number of entries and accuracy.

3.1 Dictionary-based approach

Due to the availability of many sentiment lexical resources (i.e., lexicons and dictionaries) in the English language, a number of researchers have been adopted based on these resources [4]. One of the most important methods that benefited from previous lexicons is the translation [45]. Due to the rapid development of machine translation through sites such as Google.com, Bing.com and others [45], most researchers used different translation methods to build non-English sentiment lexicons. In order to overcome the shortcomings observed in automated translation systems, researchers have opted to use multiple translations of more than one language at the same time [4, 43]. Besides the machine translation of English sentiment lexicons, other methods have been used based on existing English lexicons such as transfer learning, relationship-based approaches and merge-based approaches. Table 4 shows a survey of the studies that have built sentiment lexicons for non-English by the dictionary-based approach.

3.1.1 Translation-based approach

In the translation approach, the language which has many dependable resources (i.e., lexicons) is called the source language (e.g., English), and the lacking resource language is called the target language. The target language will identify the sentiment polarities of texts using the

Table 2 English lexicons used to produce non-English sentiment lexicons

Sentiment lexicon	References	Domain	Entry size	Polarity	Description	License
SentiWordNet[a]	[29, 30]	General domain	117,658 Synsets	Positive, negative, objective	SentiWordNet is a lexical resource publicly available for research purposes. It consists of an annotation of the WordNet by indicating the degree of each term by using numerical scores ranging from 0.0 to 1.0, which indicate the polarity of the word (positive, negative and objective). Four different versions of SentiWordNet have been published: 1.0, 1.1 [29], 2.0 and 3.0 [30]	Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license
SenticNet[b]	[14]	General domain	14,244 Common sense concepts	Positive, negative	SenticNet 2.0 is a publicly available affective commonsense resource for sentic computing. It consists of 14,244 common sense knowledge concepts [28] used for concept-level sentiment analysis. SenticNet was developed through the ensemble application of graph-mining and dimensionality-reduction techniques over multiple commonsense knowledge bases [31]	MIT License

Table 2 continued

Sentiment lexicon	References	Domain	Entry size	Polarity	Description	License
Opinion Finder[c]	[15]	General domain	6856 unique entries	Negative, Neutral, Positive	OpinionFinder system included a sentiment lexicon that is available for separate download. The lexicon composed of manually developed resources with entries extracted from corpora. OpinionFinder lexicon consists of 6856 unique entries associated with their polarity value [35]	GNU General Public License
Bing Liu's Opinion Lexicon[d]	[16]	General domain	6789 Words	− 1, 0, 1	Opinion Lexicon is a manually created lexicon by extracting the polarity words from customer reviews. The lexicon includes 4783 negative words and 2006 positive words	Free
MPQA Subjectivity Lexicon[e]	[17]	General domain	8221	Strong or weak Positive or negative	MPQA Lexicon is a manually created lexicon, consisting of 8221 words with their subjectivities (strong or weak), polarities and POS tags	GNU General Public License

Table 2 continued

Sentiment lexicon	References	Domain	Entry size	Polarity	Description	License
Harvard General Inquirer[f]	[36]	General domain	3206	Positive, negative	A manually created lexicon consisting of 3206 entries divided into 915 positive words and 2291 negative words. These marked words are divided into 182 categories such as positive, negative, strong, weak, active	Available for research purposes
AFINN[g]	[37]	General domain	2477	− 1, 0, 1	A manually collected list of English polarity words and phrases rated between + 5 (very positive) and − 5 (very negative). The first version, AFINN-96, contained 1468 unique words whereas the newest version has 2477 words	Open Database License (ODbL) v1.0
SentiFul[h]	[38]	General domain	2253	Positivity and negativity score; polarity weight	An automatically generated sentiment lexicon using the WordNet-Affect database. SentiFul has been further extended using SentiWordNet and polysemy	Available for research purposes

https://sentiwordnet.isti.cnr.it/
http://sentic.net/downloads/
http://mpqa.cs.pitt.edu/lexicons/
http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
http://www.wjh.harvard.edu/~inquirer/
https://github.com/fnielsen/afinn
https://sites.google.com/site/alenaneviarouskaya/research-1/sentiful

Table 3 Summary of the approaches used to build sentiment lexicons for non-English languages

Resources	Approaches	Overview	References
Dictionaries, Lexicons	Translation-based	This approach relies on translating an existing sentiment lexicon into a target language. Usually, machine translation or bilingual dictionaries are used	[7, 18, 21, 35, 40–53]
	Relationship-based	This approach starts with a small group of core words (seeds) that expand by using the semantic relations between words (i.e., synonyms and antonyms) in an existing dictionary	[22, 54–62]
	Merge-based	This approach uses to create large sentiment lexicons by combining predefined lexicons. It is useful in increasing the coverage and expansion of the lexicons	[63–67]
Corpus	Frequency-based	Statistical standards are used to calculate words frequency in a given polarity. This approach assumes that positive words appear together with positive words and vice versa	[41, 44, 68–70]
	Graph-based	This approach uses semantic relations between words in a large corpus to find new words related to predefined words (seeds)	[19, 23, 71]
Human	Crowdsourcing	The lexicons are built by encouraging people to answer questions or a puzzle. People select words from a text and label them with polarities using crowdsourcing and game with a purpose	[20, 72–74]
	Manual	The lexicons are created manually by researchers or linguists	[75, 76]

existing resources in the source language [4]. In general, the translated sentiment lexicon is built in three steps as shown in Fig. 2. First, translating the source language lexicon using machine translation tools, in which simple substitution of words takes place from one language to another. This is then followed by POS tagging before the target language lexicon is cleaned and filtered to remove duplicates and non-translated words [46, 58, 68]. Most of the studies were found to perform the filtering manually, probably as it is easier to find non-translated or duplicate words using sorting approaches [46, 58, 68].

One of the first studies that used machine translation to construct a sentiment lexicon for a non-English language was conducted by Yao et al. [42], who proposed an automatic translation method for building a Chinese sentiment lexicon. They used an electronic dictionary named StarDict[1] to translate Chinese words into English, followed by parsing to generate sequences of English words. These words were then used to determine the sentiment score

[1] http://stardict.sourceforge.net.

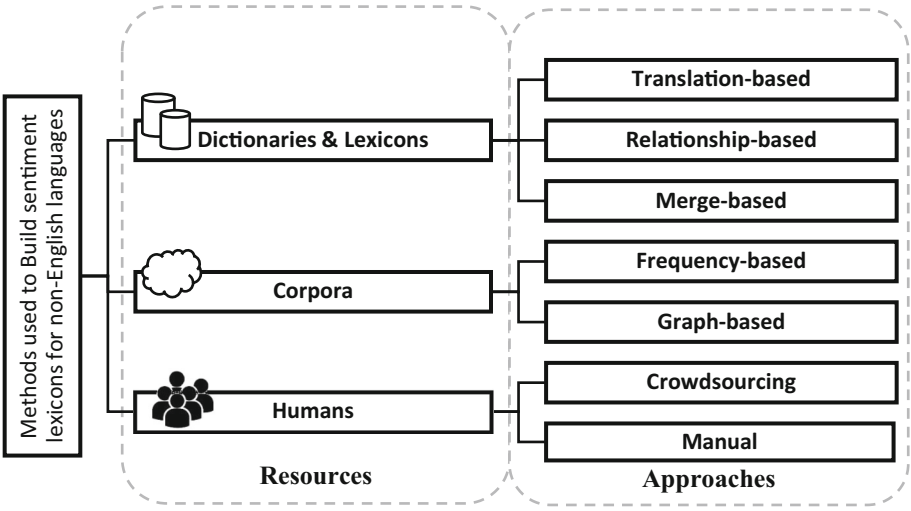

Fig. 1 The taxonomy of the approaches used to build sentiment lexicons for non-English languages

for the specific words. Likewise, Mihalcea et al. [7] used the same approach for the Romanian language where they used a bilingual dictionary to translate current English lexicon to the target language (i.e., Romanian). The Opinion-Finder [94] was used as a sentiment lexicon resource, and they translated it to the Romanian language. The authors used two bilingual dictionaries to perform the translation. One of them is an English-Romanian dictionary, and the second dictionary was obtained from the Universal Dictionary download site.[2] After translating the English obtained resources, they built a Romanian sentiment lexicon consisting of 4983 entries. Then, they built a rule-based subjectivity classifier using the new lexicon.

Steinberger et al. [43], on the other hand, proposed a triangulation method (i.e., translating two languages to produce a new lexicon in a third language) using the semiautomatic approach. In other words, the authors used machine translation to translate the sentiment lexicons in English and Spanish to Arabic, Czech, French, German, Italian, and Russian, followed by filtering and expanding the new lexicon manually. In [21], the researchers presented an idea that is not very different from previous ideas. The approach is also based on machine translation to build and elaborate a new French sentiment and emotion lexicon. Their method was based on the semiautomatic translation, where they used English NRC Word Emotion Association Lexicon [27] as a resource lexicon. They called their lexicon as (FEEL). The work was done in two stages. The first stage was conducting an online translation on existing English lexicon to create a first version of their French Lexicon. Then, they validated their automatically obtained French lexicon manually by a human professional translator. Researchers concluded that it could rely on the online translators to inexpensively extract resources using appropriate thresholds and heuristics [21]. In the Norwegian language, Hammer et al. [40] created and evaluated a large set of sentiment lexicons by using a machine translation tool (Google translate[3]) to translate AFINN English sentiment lexicon [37] from English to Norwegian. They denote this sentiment lexicon as AFINN in their work. Moreover, they generated one more lexicon to correct some obtained errors from the machine

[2] http://www.dicts.info/uddl.php.

[3] translate.google.com.

Table 4 Survey of the papers that have built sentiment lexicons for non-English (Dictionary-based approach)

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
Dictionary-based	Translation-based	French	2016	[21]	NRC-EmoLex	[27]	Six online translators	General	14,127	SVM	74.3	74.4	72.8
		Norwegian	2014	[40]	AFINN	[37]	Google translate	General	2161	N/A	N/A	N/A	N/A
		Arabic (AraSenti-Trans)	2016	[41]	Opinion Lexicon MPQA	[16] [17]	MADAMIRA tool [77]	Twitter	N/A	N/A	78.5	78.1	76.3
		Chinese	2006	[42]	10 bilingual lexicons from StarDict[a]	–	Bilingual Translator,	General	4120	SVM	83.3	91.4	N/A
										DT	84.1	92.8	N/A
		Romanian	2007	[7]	Opinion-Finder	[15]	Bilingual Translator	General	4983	LB	62.6	33.5	43.7
		Multi-languages	2012	[43]	MicroWNOp and JRC Tonality	[78] [79]	Triangulation	General	About 2000 per language	LB	N/A	N/A	N/A
		Arabic	2011	[18]	SentiStrength	[80]	Manually translation	Education, Politics, Sports	8793	LB	81.3	81.7	82.7
		German (SentiWS)	2010	[44]	General Inquirer	[36]	Google translate	General	N/A	LB	96	74	84
		Spanish	2012	[35]	Opinion-Finder SentiWordNet WordNet	[15] [30]	multilingual sense-level aligned WordNet structure	General	1347	SVM	64.6	82.4	72.4
										Manual	91.8	88.2	90.0
		German	2008	[45]	SentiWordNet	[29, 30]	Translator	General	N/A	LB	66	N/A	N/A

Table 4 continued

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
		Romanian Spanish	2013	[46]	OpinionFinder	[15]	Ectaco online dictionary [b]	General	1580 2009	SVM	67.7 66.9	38.1 50.5	48.9 57.6
		Korean	2010	[47]	OpinionFinder	[15]	Multilingual translator	General	3808	LB	59.4	71.0	64.7
		Chinese							3980		58.4	82.3	68.2
		Japanese							3027		56.9	92.4	70.4
		Italian	2013	[48]	SentiWordNet	[30]	Transfer learning	Twitter	N/A	N/A	55	N/A	N/A
		Singlish (Singaporean English)	2016	[49]	English-Malay lexicon Many online resources	[81]	Matching English polarity list with Singlish list	Twitter	2666	Hybrid	N/A	N/A	77
		Spanish (SEL)	2012	[50]	SentiWordNet	[30]	Maria Moliner dictionary [82]	Twitter	2036	NB DT SVM	78.2 83.6 85.8	N/A N/A N/A	N/A N/A N/A
		German	2006	[51]	WordNet	[83]	Translator	German Emails	3871	LB	N/A	N/A	N/A
		French, Italian, Spanish and German	2016	[53]	MPQA Opinion Lexicon General Inquirer NRC Emotion Lexicon	[17] [16] [36] [84]	Transfer learning and translating	Twitter	N/A	SVM	N/A	N/A	61.7
		Bengali	2010	[52]	SentiWordNet	[30]	English-Bengali Dictionary	General	35,805	LB	74.6	80.4	N/A

Table 4 continued

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
	Relationship-based	Arabic Hindi	2011	[54]	WordNet Arabic WordNet Hindi WordNet General Inquirer	[30] [85] [86] [36]	Random Walk	General	N/A	SO-PMI	83 93	N/A	N/A
		Persian	2014	[61]	WordNet 3.0 FarsNet 1.0	[83] [87]	Random Walk	General	4941	N/A	80.6	80.5	N/A
		Swedish	2010	[55]	People's Dictionary of Synonyms	[88]	Random Walk	General	1349	N/A	N/A	N/A	N/A
		Romanian	2008	[56]	Romanian online dictionary[c]	–	Bootstrapping Method	General	4000	LB	62.8	69.9	66.2
		Hindi French	2009	[57]	Hindi WordNet[d] OpenOffice thesaurus[e]	–	Graph-based label propagation	General	N/A	LP	90.9 73.6	95.1 93.6	93 82.5
		Arabic	2014	[58]	Arabic WordNet	[85]	Semi-supervised learning	General	7576	NB SVM	94 73	91 65	N/A N/A
		Malay	2016	[62]	WordNet	[30]	Graph-based label propagation	General	4206	NB	~ 64	N/A	N/A
		Hindi	2012	[59]	English-Hindi WordNet linking SentiWordNet	[89] [30]	Graph-based	General	8936	Human Judgment	~ 79	N/A	N/A
		Chinese	2009	[60]	HowNet semantic lexicon[f]		Semantic similarity	Hotel	5573	SVM	82.1	N/A	N/A

Table 4 continued

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
		Multi-languages	2014	[81]	Wiktionary[g]		Knowledge graph propagation	General	Based on the language	N/A	N/A	N/A	N/A
		Swedish	2016	[22]	SALDO	[90]	Semantic Relations		2127		85	N/A	N/A
	Merge-based approach	Arabic (ArSenL)	2014	[63]	SAMA SentiWordNet Arabic WordNet	[91] [29, 30] [85]	Merged existing sentiment lexicons	General	28,812	SVM	58.3	95.1	72.3
		Hindi (H-SWN)	2010	[64]	English-Hindi WordNet SentiWordNet	[89] [30]	Matching two lexical resource	General	16,253	SVM	60.3	N/A	N/A
		Arabic (SANA)	2014	[65]	SIFAAT and HUDA SentiWordNet General Inquirer	[92] [30] [36]	Merged existing sentiment lexicons, translating English lexicons	General	224,564	N/A	N/A	N/A	N/A

Table 4 continued

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
		Italian	2016	[67]	AFINN, Opinion Lexicon, SentiWordNet	[37] [16] [30]	Merged existing sentiment lexicons, translating English lexicons	General	15,412	SVM	87	66	75
		Arabic (SLSA)	2015	[66]	AraMorph SentiWordNet	[93] [30]	Own linking algorithm	General	35,000	SVM	67	66.6	68.6

The columns: Precision, Recall, F-measure, are the results of the evaluation made by the researchers on their own data

SVM support vector machines, *NB* Naïve Bayes, *DT* decision tree algorithm, *LB* lexicon-based, *LP* Label Propagation test, *CS* crowdsourcing

[a] http://goldendict.org/dictionaries.php

[b] http://www.ectaco.co.uk/free-online-dictionaries/

[c] http://www.dexonline.ro

[d] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[e] http://www.keenage.com/

[f] http://www.openoffice.org

[g] https://www.wiktionary.org/

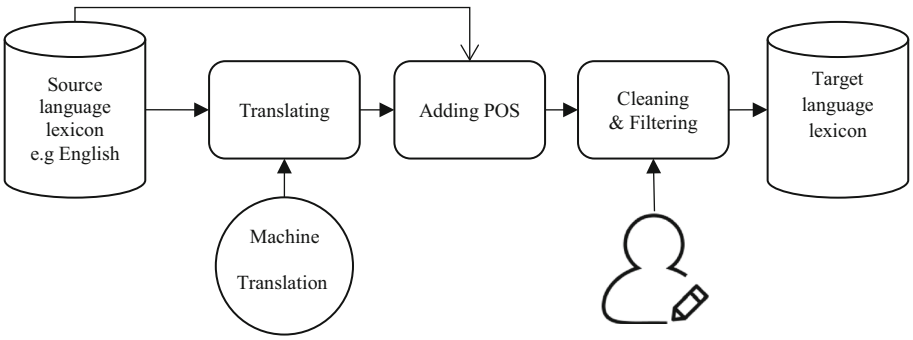

Fig. 2 The general steps of translating-based approach

translation by using a manual check. Thus, it appears clear that translation-based approaches depends on the availability of the translation engines for the required (i.e., target) languages [95].

Kim and Hovy [51] applied a German sentiment system by translating WordNet to German. They built a lexicon-based system to analyzing German emails using their German sentiment lexicon. Similarly, Denecke [45] used SentiWordNet as a lexical resource to detect the polarity of a German document within a multilingual framework by translating the English source into German.

Al-Twairesh et al. [41] generated large-scale Twitter sentiment lexicons for Arabic called AraSenti-Trans using the MADAMIRA tool [77] that identified Arabic words in tweets and removed tweets that contained non-Arabic words and dialects. After pre-processing, they used Bing Liu's Opinion Lexicon [16] and the MPQA lexicon [17] as sentiment orientation resources. They used the English gloss that was provided by MADAMIRA to find words polarity by comparing with the Liu lexicon and MPQA lexicon.

OpinionFinder [15] was used by Kim et al. [47] to build multilanguage sentiment lexicons for three languages. They used machine translation to translate OpinionFinder lexicon to Korean, Chinese and Japanese. Their approach produced Korean, Chinese and Japanese lexicons containing (3808, 3980 and 3027 entries) respectively. Similarly, Perez-Rosas et al. [35] used OpinionFinder [15] along with SentiWordNet [30] as English electronic resources to transfer the sentiment scores to Spanish. Likewise, Basile and Nissim [48] used SentiWordNet and MultiWordNet to transfer the word's polarity from English SentiWordNet to Italian sentiment lexicon. Another transfer technique was used by Das and Bandyopadhyay [52]. They used an English-Bengali Dictionary to apply a word level lexical transfer technique to each entry in English SentiWordNet. The result was a Bengali SentiWordNet with 35,805 Bengali entries.

Some studies have presented hybrid methods (i.e., using translation method with other methods), but they are mainly based on translation. A study by Sidorov et al. [50] built Spanish emotion lexicon called SEL and it contained 2036 words. The lexicon was built in three stages. First, the lexicon was based on the selection and automatic translation of the words from English SentiWordNet [30] to Spanish. Then, the developers used Maria Moliner dictionary [82] to check the obtained words and keep the words that had a meaning related to the basic emotions: joy, anger, fear, sadness, surprise and disgust. Finally, they asked 19 annotators to evaluate the association of the words with the emotions. The annotators put the scales such as null, low, medium and high for each entry [50].

General Inquirer lexicon [36] was used as a source for building German sentiment lexicon as in Remus et al. [44] work. They used Google translator to translate General Inquirer lexicon into German. In the revision phase, they manually removed any word without prior polarity. Banea et al. [46] presented sentiment lexicons for Romanian and Spanish languages. Their approach based on automatically translating a source language lexicon into Romanian and Spanish by utilizing multilingual dictionary. After translating, they expanded the lexicons using the bootstrapping process (see Sect. 3.1.2 for further details on bootstrapping).

Instead of using machine translation alone, the manual translation (i.e., translating by human translators) was used as well, such as the work provided by El-Halees [18]. He manually translated an English lexicon to the Arabic language. The Arabic sentiment lexicon was built by using two resources: the SentiStrength project [80] and an online dictionary. After the translation process, the initial list was manually filtered. Then, the same strength that was used in SentiStrength was used in the Arabic list. Finally, he used an online dictionary to add other common Arabic words; some of them were synonyms and the others were significant words added manually. Lo et al. [49] also manually constructed a Singlish (Singaporean English) sentiment lexicon. Singlish is a localized form of English that includes elements of various languages. They combined several Internet resources. The used resources included Coxford Singlish Dictionary,[4] Singlish and Singapore English[5] and Wikipedia Singlish vocabulary.[6] For English, they used many online sources such as the positive and negative lists of a Twitter sentiment analysis, and a set of positive vocabulary word lists.[7] Then, to determine the polarity of words a Malay-English sentiment lexicon [81] was used. The final list contained 2666 entries of Singlish terms.

3.1.2 Relationship-based approach

The relationship-based approach starts with a small group of core words (seeds) that expand by using the semantic relations between words in an existing dictionary or lexicon [24].

Mahyoub et al. [58] presented an algorithm that assigns sentiment scores to the entries found in the Arabic WordNet to create an Arabic sentiment lexicon. After preparing the seed list of positive and negative words (i.e., a total of 14 words), a semi-supervised learning algorithm to increase the number of entries in the Arabic WordNet was applied by exploiting the synset relations. This was accomplished by randomly selecting and including new words from the synsets that were missing in the seed list. Similarly, Nusko et al. [22] presented a method to build a sentiment lexicon for the Swedish language. A small group of core words (seeds) were expanded by using the semantic relations between words in SALDO [90]. SALDO is a lexical resource of modern Swedish. Rosell and Kann [55] also built a Swedish sentiment lexicon by using another method that was based on random walks algorithms. They started with a small group of words (seeds) taken from People's Dictionary of Synonyms [88] and used graph method to calculate distances between words. The random walk has been applied in their work to decide the next node and length of each edge. For each word, they repeated the method 10 times and they calculated the standard deviations and the mean values for each word.

[4] http://www.talkingcock.com/html/lexec.php.

[5] http://www.singlishdictionary.com/.

[6] https://en.wikipedia.org/wiki/Singlish_vocabulary.

[7] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English.

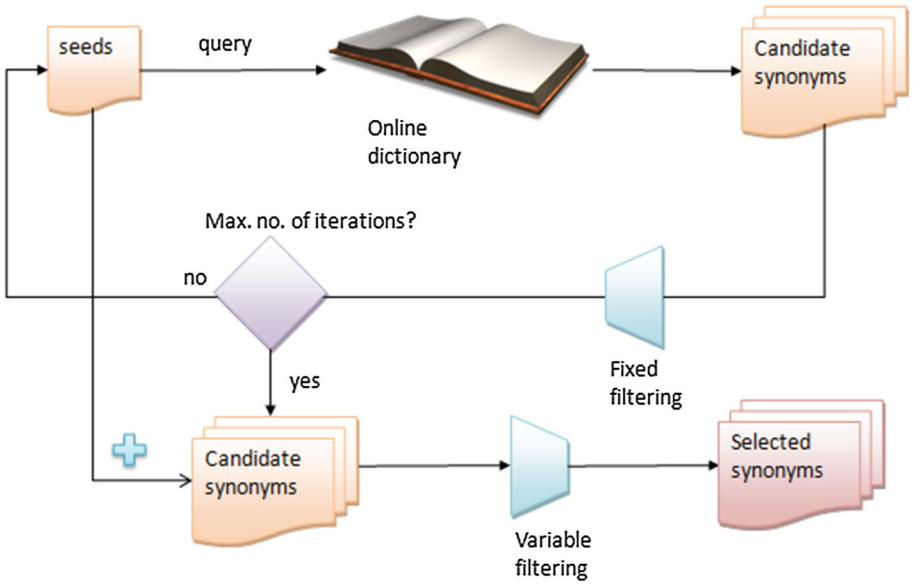

Fig. 3 Bootstrapping process by Banea et al. [56]

Hassan et al. [54] built multilingual lexicons in two languages, Arabic and Hindi. The general goal of their work was to extract the semantic orientation of new words. They created a multilingual network of words where the words will connect if they are semantically related. For example, the authors used Wordnet [83] as a source of synonyms and hypernyms for linking English words in the network (i.e., English–English). By way of an example, *color* is a hypernym for *red* while *carmine* and *sanguine* belong to the same synset *red;* hence, they are considered synonymous. Similarly, Arabic WordNet (AWN) [85, 96] and Hindi WordNet [86] were used for the Foreign–Foreign word connections in the network, whereas an English-to-foreign dictionary was used to generate the English-Foreign word connections.

Banea et al. [56] introduced a bootstrapping method to build a sentiment lexicon by generating rule-based classifiers for languages with scarce resources starting with manually picked seeds. The bootstrapping method in an online dictionary was used to extract new subjective candidates. For each seed word, a query is made in an online dictionary. From the results, a list of related words is selected and added to the list of candidates. The candidate words are filtered based on their similarity with the original seed. Then, continue to the next iteration until a maximum number of iterations is reached. The new subjective words were ranked based on the Latent Semantic Analysis (LSA) similarity measure, and the top entries were used to build a sentiment classifier [56]. Figure 3 illustrates the bootstrapping process as described in Banea et al. [56] work. This method is useful, but its problem is that it requires a synonyms dictionary for the target language.

Another graph-based framework used by Rao and Ravichandran [57] applied a graph-based framework on WordNet lexical resource to construct sentiment lexicons for both Hindi and French. They improved the label propagation results by using synonymy and hypernymy relationships (i.e., semantic relation). Hindi WordNet was used for Hindi, and OpenOffice thesaurus for French. Similarly, Bakliwal et al. [59] expanded a Hindi lexicon using a

graph-based WordNet method by expanding initial seed lexicon with synonym and antonym relations. In Chinese language, Zhu et al. [60] constructed Chinese sentiment lexicon based on HowNet semantic lexicon on the semantic similarity of Chinese words.

Additionally, many researchers implemented filtering which is accomplished by calculating similarity measures in the graph to remove noise from the lexicons. Common techniques used include Pointwise Mutual Information (PMI) [54, 97] or Latent Semantic Analysis (LSA) [56, 98]. For instance, Banea et al. [56] calculated the similarity measures between seed and candidate words in order to choose the candidates in the next iteration, with results indicating LSA to be more efficient than PMI (i.e., faster and requires less training data).

3.1.3 Merge-based approach

The main idea behind the merge-based approach is to create large sentiment lexicons by combining predefined lexicons in order to increase accuracy. This is especially useful for languages in which lexical resources are lacking such as Arabic and Hindi. The merging may take various forms, such as combining several lexicons in the same language together, or by translating several lexicons before the merge. For example, Badaro et al. [63] merged four existing sentiment lexicons to produce new Arabic sentiment lexicon called (ArSenL). The sentiment lexicons that they relied on were: Standard Arabic Morphological Analyzer [91], English WordNet, English SentiWordNet [29, 30] and Arabic WordNet [85]. Likewise, Joshi et al. [64] developed a sentiment resource for Hindi known as Hindi-SentiWordNet lexicon or H-SWN exploiting two existing resources namely English-Hindi WordNet linking [89] and English SentiWordNet [30]. The basic premise was to keep the words of the Hindi language unchanged. Thus, if a word is found in English in SentiWordNet, the algorithm searches for the corresponding word in Hindi WordNet, and the process is repeated until the corresponding words are found and added to the lexicon. The result is Hindi sentiment lexicon with sentiment scores associated with synsets. Abdul-Mageed and Diab [65] presented an Arabic sentiment lexicon called SANA. SANA is a large-scale Arabic lexicon for standard Arabic and some Arabic dialects developed both manually and automatically. The authors leveraged two existing Arabic lexicon namely SIFAAT and HUDA [92]. SIFAAT means "adjectives" in Arabic, which is composed of 3325 Arabic adjectives. The other one was HUDA lexicon, which extracted from an Egyptian Arabic chat data set. A statistical method based on pointwise mutual information (PMI) and machine translation to add extra words from English sentiment lexicons such as SentiWordNet [30], General Inquirer [36] was used where the total words collected were 224,564. Eskander and Rambow [66] constructed an Arabic sentiment lexicon called SLSA by linking the Arabic morphological analyzer lexicon (AraMorph) [93] with the SentiWordNet English lexicon [30]. When linking the resources, the sentiment scores in SentiWordNet are applied to the entries of AraMorph to generate the new sentiment lexicon.

Buscaldi and Hernandez-Farias [67] used a hybrid method to build sentiment lexicon for the Italian language. The method starts by creating a set of polarity words from the Bing Liu's Opinion Lexicon [16], AFINN [37], and SentiWordNet [30] lexicons. Their method consists of several steps beginning with the translation and merging of three dictionaries before expanding with the WordNet synonyms of words.

3.1.4 Dictionary-based approaches' limitations

When using dictionary-based approaches, several limitations emerged as follows;

- The sentiment orientation of words and the sentiment lexicons that are built by those methods are general-domain lexicons and may appear less accurate when used with specific domains.
- Sentiment lexicons do not contain many words or shortcuts that are used on social networking sites. Therefore, they cannot handle different dialects and informal or slang words, because they do not exist in dictionaries [39].
- In machine translation, several errors may arise due to cultural differences about the sentiment orientations of words (i.e., a word may be positive in one language and negative in another and vice versa) [7, 35].
- Many words are lost in translation; where they appear in the same translation and are considered duplicated in the new lexicon and automatically deleted [74]. This happens because these translators rely on the most common words. Therefore, a number of synonyms may be translated as the same word, because the most common word is used for of those synonyms. This creates a loss in the new lexicon of the target language [5, 7].

3.2 Corpus-based approach

Corpus is a large collection of computer-readable written texts, whether they are comments, documents, reviews, etc., offering a rich variety of words and structures that can be relied upon to analyze the languages [99]. Annotated corpora were used not only to build machine learning-based systems [1] but also to construct the sentiment lexicons through two types of methods: statistical methods and semantic relations methods. The statistical methods use large corpora with statistical equations to obtain polarity words in order to generate a new sentiment lexicon by calculating words frequency in particular class [100]. The second methods are used semantic relations between words in a large corpus to extract a sentiment lexicon [100]. Table 5 presents some works carried out for corpus-based approach with the identification of the languages used.

3.2.1 Frequency-based approach

Statistical equations and functions are used to calculate word frequency in a given polarity. This approach assumes that positive words appear together with positive words and vice versa. Remus et al. [44] used the co-occurrence analysis of product reviews that consisted of 5100 positive reviews and 5100 negative reviews where users determined the polarity of each comment between one and five. The results were lists of words that often appears together with one of the polarities (positive or negative).

AraSenti-PMI is an Arabic sentiment lexicon that was built using pointwise mutual information (PMI) measure in a dataset of tweets by Al-Twairesh et al. [41]. They used PMI measure to distinguish the association between words in the corpus to be classified into positive or negative words. The PMI measure was first used in sentiment analysis by Turney [103]. The PMI between two words, word1 and word2, is defined as follows.

$$\text{PMI}\ (\text{word}_1, \text{word}_2) = \log_2 \left\{ \frac{p(\text{word}_1\ \&\ \text{word}_2)}{p(\text{word}_1)p(\text{word}_2)} \right\} \tag{1}$$

The PMI measure was also used by Turney and Littman [104] to determine the polarity of a specific word by including the sentiment orientation (SO-PMI). As shown in Eq. (2)

Table 5 Survey of the papers that have built sentiment lexicons for non-English (Corpus-based)

Approach	Method	Languages (Lexicon Name)	Year	References	Data sources	Data sources Ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
Corpora-based	Frequency-based	Arabic (AraSenti-PMI)	2016	[41]	Tweets from Twitter	N/A	PMI	Twitter	N/A	N/A	90.1	89.2	89.5
		German	2010	[44]	10,200 product reviews	N/A	PMI	Business	N/A	N/A	96	74	84
		Persian	2017	[69]	7500 reviews	N/A	Semantic Orienta-tion	Business	3705	N/A	80	80	N/A
		Chinese	2013	[70]	Hotel review from lvping.com	N/A	Semantic Orientation-PMI	Business	N/A	LB NB	92.4	N/A	N/A
		Italian	2015	[101]	Corpus	N/A	PMI	General	N/A	CS	73	N/A	N/A
		Arabic (UCBSL)	2019	[102]	Corpus	N/A	Statistical method	News	6356	LB	0.78	0.72	0.74
		Hindi (HMDSAD)	2015	[68]	product reviews from	Amazon[a]	PMI	Business	N/A	N/A	N/A	N/A	N/A
	Graph-based	Arabic	2010	[71]	large web corpus from the internet	–	similarity graph	Business	1600	N/A	75–88	60–80	N/A

Table 5 continued

Approach	Method	Languages (Lexicon Name)	Year	References	Data sources	Data sources Ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
		Chinese	2015	[19]	30 million Chinese microblogs	Weibo API[b]	mutual reinforcement random walk model	Microblog	12,799	LB	52	74	61
		Polish	2014	[23]	3222 web documents		Random Walk	General	N/A	N/A	69.7	N/A	N/A

The columns: Precision, Recall, *F*-measure, are the results of the evaluation made by the researchers on their own data

[a] http://Amazon.com

[b] http://weibo.com

below the SO-PMI is calculated by basically subtracting the PMI between a word and a set of negative words (i.e., negative paradigms) from the PMI of the positive paradigms [4, 104]:

$$\text{SO - PMI (word)} = \text{PMI (word, \{positive paradigms\})} - \text{PMI (word, \{negative paradigms\})} \quad (2)$$

Jha et al. [68] created a Hindi sentiment lexicon called HMDSAD that was based on words co-occurring together frequently in a review. To calculate the relationships between the words, they used Pointwise Mutual Information (PMI). The data resource was product reviews from Amazon.com after translating it to Hindi by a Hindi translator.

3.2.2 Graph-based approach

In [19], the authors utilized a massive purified microblog dataset as training corpus to build a Chinese sentiment lexicon using emoticons to extract the polarity words in a microblog. The research found that an emoticon expresses more obvious emotions if it often co-occurs with sentiment words and other important emoticons. Thus, they observed that the positive words frequently occur with positive emoticons. Similarly, the negative words often appear with negative emoticons. They integrated the emoticons and candidate sentiment words to build a graph and ran a random walk algorithm to extract the opinion words in order to build a sentiment lexicon.

Elhawary and Elfeky [71], on the other hand, produced a similarity graph to build an Arabic sentiment lexicon that clusters all the words/phrases of a certain language. If two words have an edge, they are similar and similar words have the same polarity of sentiment, or they have the same meaning. A label propagation on similarity graph was performed on a seed list of 1600 words and 1500 features such as the frequency of keywords in the document and frequency of bolded keywords. Their lexicon consisted of two columns, one contained the word/phrase and the second column represented the score of the word. For pruning purpose, filtering rules were applied to avoid both the sparseness of the data and rubbish nodes.

Haniewicz et al. [23] attempted on building a Polish sentiment lexicon by applying the Random Walk Approach. They gathered 356,275 reviews from several goods and services websites with each review having a rating score between 1 and 5. The authors designed a semantic network to store each term in the review based on the type of relation (i.e., synonymous, hypernymous or homonymous), and their respective sentiment scores (i.e., positive, negative and neutral) in a given domain. Finally, the Polish sentiment lexicon has about 27,000 words with the value of sentiments for a given domain.

3.2.3 Corpus-based approaches' limitations

When using corpus-based approaches, several limitations emerged as follows;

- The lack of data pre-processing tools in many languages makes it difficult and complex to rely on the corpus to build lexicons
- There is a lack of adequate corpus online; especially for languages with fewer resources
- Constructing sentiment lexicons by analyzing the corpus requires a large corpus volume so that an acceptable accuracy is achieved
- The obtained lexicon does not contain many words and often only serves a particular domain. It can therefore not be relied upon to analyze another domain
- Finally, some methods depend on an annotated corpus. This requires additional data annotation before analysis can begin [105].

3.3 Human-based computing approach

This approach is to encourage people to answer questions or a puzzle in order to benefit from these answers in the construction of the sentiment lexicons. Words are selected from a text and labeled with polarities [72]. The systems are often developed using crowdsourcing platforms such as Amazon Mechanical Turk[8] [27], by building games with a purpose [72], or directly by human experts [8]. Table 6 presents works carried out for human-based approach with identification of the languages used.

3.3.1 Crowdsourcing

Many researchers proposed crowdsourcing games to construct sentiment lexicons for resource-scarce languages [72]. Lafourcade et al. [20] developed an online game with a purpose (GWAP) of asking the player to indicate the polarity and the emotion of the displayed words and terms. The players can choose the right polarity for the displayed words by pressing one of the three emoticons that represent the polarities: positive, negative and neutral. In their next project called *Emot*[9] [106], the authors improved the game by offering the player to associate one or several emotions to a given word, either by choosing one among the displayed emotions (e.g., fear, joy, love, sadness, etc.) or by entering some other emotions via a text field, if none of the presented emotions suits the player. When the researchers designed the game, they adopted the principle of simplicity in play and judgment based on the majority opinion. Hong et al. [72] developed a language-independent crowdsourcing game to build a Korean sentiment lexicon. The goal of their project was to design a game with a purpose that produces sentiment lexicons for resource-scarce languages. They called it *Tower of Babel*. Unlike previous methods, Tower of Babel required a lot of volunteers and amateur to participate in the game to build the sentiment lexicon. Therefore, no need to use a previous thesaurus or provide linguistic expertise. They designed the game like Tetris where the pieces are accumulated on top of each other. The game was a collaborative game in which a pair of volunteers agrees to make sentiment classifications on particular terms whereby the volunteers are rewarded for making a matching classification with the partner. Another idea based on teamwork was presented by Al-Subaihin et al. [73]. They proposed a game to create Arabic sentiment lexicons by encouraging players to select words from the text and label them with polarities. The game starts with two teams of two players facing each other in three rounds. They used Qaym.com as a resource for the sentences that shown to every team. The winner team is the one whose members agree on the words and feelings they have chosen. Scharl et al. [74] presented crowdsourcing games with a purpose. They designed a game called Sentiment Quiz. The idea was that a number of players from different countries speaking different languages evaluate the words of their language. The results show that more than 3500 users added approximately 325,000 evaluations in various languages such as (English, Portuguese, French, Italian, Russian, German and Spanish). The next stage was sentiment lexicons extension that was done by means of a bootstrapping process (see Sect. 3.1.2 for further details on bootstrapping). The results were satisfactory; however, the challenge here is the difficulty in convincing the required number of players or volunteers to participate in the game.

[8] https://www.mturk.com/.

[9] http://www.jeuxdemots.org/emot.php.

Table 6 Survey of the papers that have built sentiment lexicons for non-English (human-based)

Approach	Method	Languages (lexicon name)	Year	References	Data sources	Data sources Ref.	Technique	Domain	Number of entries	Evaluation method	Precision	Recall	*F*-measure
Human-Based	Crowdsourcing	Korean	2013	[72]	Players	N/A	Game (Tower of Babel)	General	N/A	N/A	N/A	N/A	N/A
		French	2015	[20]	Players	N/A	Game (LikeIt)	General	385,000	N/A	N/A	N/A	N/A
		Arabic	2011	[73]	Qaym. com Players	N/A	Game	Business	N/A	N/A	N/A	N/A	N/A
		English, Portuguese, French, Italian, Russian, German and Spanish	2012	[74]	Players	N/A	Game and bootstrapping	Social media	N/A	NB	N/A	N/A	N/A
	Manual-based	Thai	2016	[75]	linguists	Pantip[a] Twitter	web-based sentiment tagging tool (SenseTag)	General Business	5120	N/A	N/A	N/A	N/A
		Arabic	2014	[76]	Manually	N/A	Manually	News	3982	N/A	N/A	N/A	N/A

The columns: Precision, Recall, *F*-measure, are the results of the evaluation made by the researchers on their own data

[a] https://pantip.com/

3.3.2 Manual-based approach

As the name implies, the lexicons are built manually by researchers or linguists/experts in this approach. For instance, Trakultaweekoon and Klaithin [75] developed a web-based sentiment tagging tool called SenseTag to annotate data more easily. This was accomplished by training the tool based on manual annotations provided by linguists who tags each word in randomly selected sentences (i.e., positive, negative, feature, and entity). Similarly, Abdul-Mageed et al. [76] manually created a sentiment lexicon consisting of 3982 adjectives. The lexicon is part of the SAMAR system developed to analyze the Arabic subjectivity and sentiments in both Modern Standard Arabic and Arabic dialects. Although deemed to be time-consuming and expensive, this approach is still used by many researchers especially those exploring sentiment analysis in languages that lack lexical resources [76].

3.3.3 Human-based approaches' limitations

Sentiment lexicons built by humans are usually more accurate than others [72]. However, the production of these lexicons is time-consuming and requires a large number of people and is costly [27]. To overcome these problems, several researchers built electronic games, such as *Tower of Babel* [72] and *Like it!* [20].

4 Challenges and open issues

This section explains common challenges in constructing sentiment lexicons for non-English languages, which include the scarcity of initial resources, the lack of pre-processing tools [41] and translation errors [7, 35]. There are also open issues for research and development that include the impact of lexicon size on the accuracy of classification, adapting sentiment lexicons to a specific domain, and use of deep learning for building sentiment lexicons.

4.1 Work on scarce resource languages

Based on our study, we note that the efforts in analyzing sentiment on scarce resource languages are predominately devoted to making use of an available lexicon for constructing polarity lexicons. However, many non-English languages suffer from a scarcity of primary sources and tools for the construction of sentiment lexicons [4, 5]. Bilingual dictionaries, annotated corpuses and/or machine translation tools should be available for the construction process of any new lexicon.

4.2 Pre-processing tools

Sentiment analysis is a high-level NLP task that relies on pre-processing tasks, such as parsing, POS tagging, stop-word removal, stemming and word segmentation [13]. Many sentiment lexicon-building methods rely on the sources of the target language. This increases the importance of the pre-processing tools for non-English languages. The primary pre-processing steps include:

- *Text normalization and cleaning* that include converting all letters and words to an appropriate format based on the language. Moreover, it also includes converting or removing

numbers, punctuations, white spaces, diacritics and stop words. Although the removal of stop words has been shown to improve sentiment analysis performance [4, 63], it has been found to be ineffective in other cases, such as in machine translation studies [26, 107].

- *Tokenization or segmentation* is used to split the given text into smaller pieces called tokens. The method varies from one language to another, depending on the properties of the language. Common technique used is to separate the text based on the white space, such as in English and Arabic while complex word segmentation algorithms and tools such as ICTCLAS and THULAC are used for languages with no white space (e.g., Chinese and Japanese). Table 7 shows some of the tools and algorithms available for non-English languages [13], in which it can be observed that tokenization remains as a core process in all the tools and algorithms.
- *Stemming and lemmatization* are the processes of extracting the root of each word to treat a group of words that are derived from the same root as synonyms. For example, the words *playing* and *played* will be reduced to *play*. However, there is a danger of over-stemming and under-stemming [108]. Over-stemming occurs when two different words are converted to the same stem (e.g., "universal" and "university" are converted to "universe"), whereas under-stemming errors occur when words of the same concept are stemmed to different roots (e.g., the words "data" and "datum" to "dat" and "datu," respectively). Although stemming is a common step in text pre-processing, it is nevertheless language dependent [109]. For example, as shown in Table 7, the majority of tools supporting the Chinese language do not have stemming. On the other hand, lemmatization takes the morphological analysis of the words into consideration [110].
- *POS tagging* aims to specify parts of speech to each word of a given text (such as adjectives, verbs, nouns) based on its context and definition. POS presents useful information in sentiment analysis as some words are ambiguous in nature, for example, the word "novel" is a neutral noun, but a positive adjective [111].

4.3 Effect of lexicon size on classification accuracy

As shown in Tables 4, 5 and 6, sentiment lexicon size did not have a significant impact on classification accuracy. Huge lexicons were therefore considered a challenge to sentiment analysis [117]. It is therefore necessary to study the accuracy of words in lexicons—not the number of words. In work presented by Badaro et al. [63], the Arabic lexicon size used was 28,812 entries, but the total precision was 58.3%. However, in the same language, Mahyoub et al. [58] constructed a general-domain lexicon that consisted of 7576 entries, with a total precision of 94%. As a result, the lexicon size is not a criterion for evaluating the lexicon accuracy. However, the lexicon should cover most of the necessary sentiment words for the classification process. Moreover, the sentiment lexicon size is also different from one language to another.

4.4 Adapting sentiment lexicons to specific domains

Lexicons extracted from general-domain lexicons are unable to deal with sentiment information from another domain [8, 118]; the reason being general-domain lexicons include formal language rather than informal expressions. Moreover, the sentiment orientations of some words vary from domain to domain. Therefore, resources expanded from those general-domain lexicons will exhibit limitations when used with non-English languages [5]. Based on Tables 4, 5 and 6, 61% of the works surveyed in our study built general-domain lexicons

Table 7 Some available NLP tools used for non-English languages

Name	Features					Languages	Developer	Programming language	License
	Lem./Stem.	Token/Seg	POS	NER	Others				
GATE	✓	✓	✓	✓	✓	Multi-language	University of Sheffield (1995)	JAVA	GNU Lesser General Public License
Stanford CoreNLP[a] [112]	✓	✓	✓	✓	✓	Multi-language	Stanford NLP Group	JAVA	GNU Lesser General Public License
spaCy[b] [113]	✓	✓	✓	✓	✓	Multi-language	Explosion AI, 2016	Python/Cython	MIT License
Natural Language Toolkit (NLTK)[c]	✓	✓	✓	✓	✓	Multi-language	The University of Pennsylvania, 2001	Python	Apache 2.0
FreeLing[d]	✓	✓	✓	✓	✓	Multi-language	TALP, Universitat Politècnica de Catalunya	C++	Affero GPL
FudanNLP[e] [114]		✓	✓	✓	✓	Chinese	Fudan University	JAVA	GNU Lesser General Public License
Apache OpenNLP	✓	✓	✓	✓	✓	Multi-language	Apache Software Foundation, 2004	JAVA	Apache License, version 2.0
FARASA[f] [115]	✓	✓	✓	✓	✓	Arabic	ALT Group	JAVA	Open source
MADAMIRA[g]	✓	✓	✓	✓	✓	Arabic and English	Diab et al. [77]	JAVA	Free for research only
ICTCLAS[h]		✓	✓			Chinese	Huaping et al. [116]	C++, Java, Python	
THULAC[i]		✓	✓			Chinese	Tsinghua University	Python	Free for research only

Table 7 continued

Name	Features					Languages	Developer	Programming language	License
	Lem./Stem.	Token/Seg	POS	NER	Others				
TextBlob	✓	✓	✓	✓	✓	Multi-language		Python	
Jieba[j]		✓				Chinese	Open source	Python	MIT License
CKIP Segmenter[k]		✓	✓		✓	Chinese	CKIP Group	Python	MIT License
Indic NLP[l]	✓	✓			✓	Indian languages	Anoop Kunchukuttan	Python	GNU General Public License

Multi-language: more than five languages. Other Features: such as parsing, n-grams chunking, co-reference resolution
Lem Lemmatization, *Stem* stemming, *Token* tokenization, *POS* part of speech tagging, *NER* named entity recognition, *Seg* segmentation
[a]https://nlp.stanford.edu/software/l
[b]https://spacy.io/
[c]http://www.nltk.org/index.html
[d]http://nlp.lsi.upc.edu/freeling/node/1
[e]https://github.com/FudanNLP/fnlp
[f]http://qatsdemo.cloudapp.net/farasa/
[g]https://camel.abudhabi.nyu.edu/madamira/
[h]http://ictclas.nlpir.org/
[i]http://thulac.thunlp.org/
[j]https://github.com/LiveMirror/jieba
[k]http://ckipsvr.iis.sinica.edu.tw/
[l]https://github.com/anoopkunchukuttan/indic_nlp_library

while 39% of them built specific domain lexicon. Accordingly, the area of building specific domain lexicons for non-English languages still require significant improvement.

4.5 Lack of evaluation benchmarks

One of the most critical challenges is the lack of evaluation benchmarks [26, 119]. As a result, the performance evaluation measures of the lexicons that are shown in Tables 4, 5 and 6 vary from one research to another. Most researchers use accuracy, precision, recall and the f-measure as evaluation measures. In addition, most of the researchers evaluated their proposed methods using their own data; hence, comparison of different methods with different datasets and settings is very difficult and biased. In addition, there are a number of papers that have not provided information on the evaluation of their work. In general, significant work is still required to improve precision levels in this area.

4.6 Use of deep learning

Deep learning is one of the machine learning fields applied to solve perceptual problems such as natural languages processing and speech recognition. The approach typically includes two steps for the text-related tasks: learning word embeddings from the text and using them to provide the document representations [26]. Researchers such as Tang et al. [120] presented a deep learning approach to build sentiment lexicon that could learn sentiment information based on distant supervision [121]. However, their approach could not infer the sentiment polarity of phrases not covered by the existing vocabulary. In [122], sentence-level sentiment polarity, context information and word-level sentiment polarity were combined to learn features of words in the corpus in order to construct a microblog-specific Chinese sentiment lexicon whereas word embeddings were used to define the association between positive sentiment and Twitter words in [123]. On the other hand, Dong and de Melo [124] developed a cross-lingual propagation algorithm that generates sentiment embedding vectors for various languages, using English as the source language. Finally, a library containing generic pre-trained word vectors was developed for 294 languages, trained on a large-scale corpus[10] [125]. Deep learning is promising, however, using the approach to generate new sentiment words or phrases from the corpus remains a challenge both in English and non-English texts [26, 126].

5 Conclusion and future work

This study provides a comprehensive review of existing research performed during the period 2006–2018, on building sentiment lexicons for non-English languages. The methods employed to construct sentiment lexicons appear in three groups each dependent on the data sources used: existing lexicons, corpus and humans. The research identifies that most researchers utilize different translation methods to build non-English lexicons. Besides automatic translation of English sentiment lexicons, other methods are used based on English lexicons such as; transfer learning, the graph-based approach and the merge-based approach. Conversely, a few studies were based on target language corpus using statistical methods to analyze the corpus and extract new polar words to build new sentiment lexicons. The

[10] https://github.com/facebookresearch/fastText.

human-based approach adopted by several researchers collected the data directly from the person or individual. Crowdsourcing and "game with a purpose," were used to encourage people to select the correct polarity for each word. However, some of the research was not reviewed as it was written in Chinese and other languages but not in English [127]. Thus, in our future work, we aspire to add some of those researches.

It is also recommended that future work should be carried out to develop new methods for building sentiment lexicons for several non-English languages such as Arabic and Hindi, preferably automatic and leveraging available resources in each language. Additionally, challenges in the construction of sentiment lexicons require further research, particularly analyzing the pre-processing tools supporting non-English languages so that effective comparisons and recommendations can be provided. Based on the research and analysis performed in this study, multilingual sentiment analysis continues to suffer from limited resources and the recognition that this is an area (domain) of significant potential, requiring immediate attention.

References

1. Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167
2. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6(12):e26752
3. Akhtar MS, Gupta D, Ekbal A, Bhattacharyya P (2017) Feature selection and ensemble construction: a two-step method for aspect based sentiment analysis. Knowl Based Syst 125:116–135
4. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, Zhou Q (2016) Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput 8(4):757–771
5. Lo SL, Cambria E, Chiong R, Cornforth D (2016) Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 28:499–527
6. Biltawi M, Etaiwi W, Tedmori S, Hudaib A, Awajan A (2016) Sentiment classification techniques for Arabic language: a survey. In: 7th international conference on information and communication systems, ICICS 2016. Institute of Electrical and Electronics Engineers Inc
7. Mihalcea R, Banea C, Wiebe JM (2007) Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th annual meeting of the association of computational linguistics
8. Deng S, Sinha AP, Zhao H (2017) Adapting sentiment lexicons to domain-specific social media texts. Decis Support Syst 94:65–76
9. Wu S, Wu F, Chang Y, Wu C, Huang Y (2019) Automatic construction of target-specific sentiment lexicon. Expert Syst Appl 116:285–298
10. Ahire S (2014) A survey of sentiment lexicons. Computer Science and Engineering IIT Bombay, Bombay
11. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
12. Montoyo A, Martínez-Barco P, Balahur A (2012) Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 53(4):675–679
13. Sun S, Luo C, Chen J (2017) A review of natural language processing techniques for opinion mining systems. Inf Fusion 36:10–25
14. Cambria E, Speer R, Havasi C, Hussain A (2010) SenticNet: a publicly available semantic resource for opinion mining. In: AAAI fall symposium: commonsense knowledge
15. Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: a system for subjectivity analysis. In: Proceedings of HLT/EMNLP on interactive demonstrations. Association for Computational Linguistics
16. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM
17. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics
18. El-Halees A (2011) Arabic opinion mining using combined classification approach. In: The international Arab conference on information technology, pp 10–13

19. Feng S, Song KS, Wang DL, Yu G (2015) A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. World Wide Web Internet Web Inf Syst 18(4):949–967
20. Lafourcade M, Joubert A, Le Brun N (2015) Collecting and evaluating lexical polarity with a game with a purpose. In: RANLP
21. Abdaoui A, Azé J, Bringay S, Poncelet P (2016) FEEL: a French expanded emotion lexicon. Lang Resour Eval 51:1–23
22. Nusko B, Tahmasebi N, Mogren O (2016) Building a sentiment lexicon for Swedish. In: Digital humanities 2016. From digitization to knowledge 2016: resources and methods for semantic processing of digital works/texts, proceedings of the workshop, 11 July 2016, Krakow, Poland. Linköping University Electronic Press
23. Haniewicz K, Kaczmarek M, Adamczyk M, Rutkowski W (2014) Polarity lexicon for the polish language: design and extension with random walk algorithm. In: Swiatek J et al (eds) International conference on systems science, ICSS 2013. Springer, Berlin, pp 173–182
24. Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Based Syst 89:14–46
25. Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 28(2):15–21
26. Giachanou A, Crestani F (2016) Like it or not: a survey of twitter sentiment analysis methods. ACM Comput Surv (CSUR) 49(2):28
27. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465
28. Cho H, Kim S, Lee J, Lee JS (2014) Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. Knowl Based Syst 71:61–71
29. Esuli A, Sebastiani F (2007) SENTIWORDNET: a high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02. http://nmis.isti.cnr.it/sebastiani/Publications/2007TR02.pdf
30. Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC
31. Poria S, Gelbukh A, Hussain A, Howard N, Das D, Bandyopadhyay S (2013) Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intell Syst 28(2):31–38
32. Hung C, Lin H-KJIIS (2013) Using objective words in SentiWordNet to improve word-of-mouth sentiment classification. IEEE Intell Syst 2:47–54
33. Plutchik R (2001) The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
34. Araujo M, Reis J, Pereira A, Benevenuto F (2016) An evaluation of machine translation for multilingual sentence-level sentiment analysis. In: Proceedings of the 31st annual ACM symposium on applied computing. ACM
35. Perez-Rosas V, Banea C, Mihalcea R (2012) Learning sentiment lexicons in Spanish. In: Lrec 2012—eighth international conference on language resources and evaluation, pp 3077–3081
36. Stone PJ, Dunphy DC, Smith MS (1966) The general inquirer: a computer approach to content analysis. M.I.T. Press, Oxford, p 651
37. Nielsen FA (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: 1st workshop on making sense of microposts 2011: big things come in small packages, #MSM 2011—co-located with the 8th extended semantic web conference, ESWC 2011. Heraklion, Crete
38. Neviarouskaya A, Prendinger H, Ishizuka M (2009) SentiFul: generating a reliable lexicon for sentiment analysis. In: 2009 3rd international conference on affective computing and intelligent interaction and workshops, ACII 2009, Amsterdam
39. Wu F, Huang Y, Song Y, Liu S (2016) Towards building a high-quality microblog-specific Chinese sentiment lexicon. Decis Support Syst 87:39–49
40. Hammer H, Bai A, Yazidi A, Engelstad P (2014) Building sentiment lexicons applying graph theory on information from three norwegian thesauruses. Norsk Informatikkonferanse (NIK)
41. Al-Twairesh N, Al-Khalifa H, Al-Salman A (2016) AraSenTi: large-scale twitter-specific arabic sentiment lexicons. In: Association for computational linguistics, pp 697–705
42. Yao J, Wu G, Liu J, Zheng Y (2006) Using bilingual lexicon to judge sentiment orientation of Chinese words. In: The sixth IEEE international conference on computer and information technology, 2006. CIT'06. IEEE
43. Steinberger J, Ebrahim M, Ehrmann M, Hurriyetoglu A, Kabadjov M, Lenkova P, Steinberger R, Tanev H, Vázquez S, Zavarella V (2012) Creating sentiment dictionaries via triangulation. Decis Support Syst 53(4):689–694

44. Remus R, Quasthoff U, Heyer G (2010) SentiWS—a publicly available German-language resource for sentiment analysis. In: LREC
45. Denecke K (2008) Using sentiwordnet for multilingual sentiment analysis. In: IEEE 24th international conference on data engineering workshop, 2008. ICDEW 2008. IEEE
46. Banea C, Mihalcea R, Wiebe J (2013) Porting multilingual subjectivity resources across languages. IEEE Trans Affect Comput 4(2):211–225
47. Kim J, Li J-J, Lee J-H (2010) Evaluating multilanguage-comparability of subjectivity analysis systems. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics
48. Basile V, Nissim M (2013) Sentiment analysis on Italian tweets. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis
49. Lo SL, Cambria E, Chiong R, Cornforth D (2016) A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. Knowl Based Syst 105:236–247
50. Sidorov G, Miranda-Jiménez S, Viveros-Jiménez F, Gelbukh A, Castro-Sánchez N, Velásquez F, Díaz-Rangel I, Suárez-Guerra S, Treviño A, Gordon J (2012) Empirical study of machine learning based approach for opinion mining in tweets. In: Mexican international conference on artificial intelligence. Springer
51. Kim S-M, Hovy E (2006) Identifying and analyzing judgment opinions. In: Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics. Association for Computational Linguistics
52. Das A, Bandyopadhyay S (2010) Sentiwordnet for bangla. Knowl Shar Event4 Task 2:1–8
53. Rouvier M, Favre B (2016) Building a robust sentiment lexicon with (almost) no resource. arXiv preprint arXiv:1612.05202
54. Hassan A, Abu-Jbara A, Jha R, Radev D (2011) Identifying the semantic orientation of foreign words. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2. Association for Computational Linguistics
55. Rosell M, Kann V (2010) Constructing a swedish general purpose polarity lexicon random walks in the people's dictionary of synonyms. SLTC 2010:19
56. Banea C, Wiebe JM, Mihalcea R (2008) A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In: Proceedings of the international conference on language resources and evaluation, LREC 2008, 26 May–1 June 2008, Marrakech, Morocco, pp 2764–2467
57. Rao D, Ravichandran D (2009) Semi-supervised polarity lexicon induction. In: Proceedings of the 12th conference of the European chapter of the association for computational linguistics. Association for Computational Linguistics
58. Mahyoub FHH, Siddiqui MA, Dahab MY (2014) Building an Arabic sentiment lexicon using semi-supervised learning. J King Saud Univ Comput Inf Sci 26(4):417–424
59. Bakliwal A, Arora P, Varma V (2012) Hindi subjective lexicon: a lexical resource for hindi polarity classification. In: Proceedings of the eight international conference on language resources and evaluation (LREC)
60. Zhu Y, Wen Z, Wang P, Peng Z (2009) A method of building Chinese basic semantic lexicon based on word similarity. In: 2009 Chinese conference on pattern recognition, CCPR 2009 and the 1st CJK joint workshop on pattern recognition, CJKPR, Nanjing
61. Dehdarbehbahani I, Shakery A, Faili H (2014) Semi-supervised word polarity identification in resource-lean languages. Neural Netw 58:50–59
62. Darwich M, Noah SAM, Omar N (2016) Automatically generating a sentiment lexicon for the Malay language. Asia Pac J Inf Technol Multimed 5(1):49–59
63. Badaro G, Baly R, Hajj H, Habash N, El-Hajj W (2014) A large scale Arabic sentiment lexicon for Arabic opinion mining. ANLP 2014:165
64. Joshi A, Balamurali A, Bhattacharyya P (2010) A fall-back strategy for sentiment analysis in hindi: a case study. In: Proceedings of the 8th ICON
65. Abdul-Mageed M, Diab MT (2014) SANA: a large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In: LREC
66. Eskander R, Rambow O (2015) SLSA: a sentiment lexicon for Standard Arabic. In: Conference on empirical methods in natural language processing, EMNLP 2015. Association for Computational Linguistics (ACL)
67. Buscaldi D, Hernandez-Farias DI (2016) IRADABE2: lexicon merging and positional features for sentiment analysis in Italian. In: CLiC-it/EVALITA
68. Jha V, Savitha R, Hebbar SS, Shenoy PD, Venugopal K (2015) Hmdsad: Hindi multi-domain sentiment aware dictionary. In: 2015 International conference on computing and network communications (CoCoNet). IEEE

69. Rashed FE, Abdolvand N (2017) A supervised method for constructing sentiment lexicon in Persian language. J Comput Robot 10(1):11–19
70. Yang AM, Lin JH, Zhou YM, Chen J (2013) Research on building a Chinese sentiment lexicon based on SO-PMI. In: Zhang J et al (eds) Information technology applications in industry, Pts 1-4. Trans Tech Publications Ltd, Stafa-Zurich, pp 1688–1693
71. Elhawary M, Elfeky M (2010) Mining Arabic business reviews. In: 2010 IEEE international conference on data mining workshops (ICDMW). IEEE
72. Hong Y, Kwak H, Baek Y, Moon S (2013) Tower of babel: a crowdsourcing game building sentiment lexicons for resource-scarce languages. In: 22nd international conference on World Wide Web, WWW 2013, Rio de Janeiro
73. Al-Subaihin, A.A., H.S. Al-Khalifa, and A.S. Al-Salman. A proposed sentiment analysis tool for modern arabic using human-based computing. in Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services. 2011. ACM
74. Scharl A, Sabou M, Gindl S, Rafelsberger W, Weichselbraun A (2012) Leveraging the wisdom of the crowds for the acquisition of multilingual language resources. In: 8th international conference on language resources and evaluation (LREC-2012), 23–25 May 2012, Istanbul, Turkey, pp 379–383
75. Trakultaweekoon K, Klaithin S (2016) SenseTag: a tagging tool for constructing Thai sentiment lexicon. In: 2016 13th international joint conference on computer science and software engineering (JCSSE). IEEE
76. Abdul-Mageed M, Diab M, Kübler S (2014) SAMAR: subjectivity and sentiment analysis for Arabic social media. Comput Speech Lang 28(1):20–37
77. Pasha A, Al-Badrashiny M, Diab MT, El Kholy A, Eskander R, Habash N, Pooleery M, Rambow O, Roth R (2014) MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: LREC
78. Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G (2007) Micro-WNOp: a gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: Language resources and linguistic theory: typology, second language acquisition, English linguistics, pp 200–210
79. Balahur A, Steinberger R, Van Der Goot E, Pouliquen B, Kabadjov M (2009) Opinion mining on newspaper quotations. In: IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies, 2009. WI-IAT'09. IEEE
80. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544–2558
81. Chen Y, Skiena S (2014) Building sentiment lexicons for all major languages. In: 52nd annual meeting of the association for computational linguistics, ACL 2014. Association for Computational Linguistics (ACL), Baltimore, MD
82. Moliner M (1984) Diccionario de uso del espanol.-v. 1–2
83. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41
84. Mohammad S, Turney P (2013) NRC emotion lexicon, in National Research Council. NRC Technical Report, Canada
85. Black W, Elkateb S, Rodriguez H, Alkhalifa M, Vossen P, Pease A, Fellbaum C (2006) Introducing the Arabic WordNet project. In: Proceedings of the third international WordNet conference
86. Narayan D, Chakrabarti D, Pande P, Bhattacharyya P (2002) An experience in building the indo WordNet—a WordNet for Hindi. In: First international conference on global WordNet, Mysore, India
87. Shamsfard M, Hesabi A, Fadaei H, Mansoory N, Famian A, Bagherbeigi S, Fekri E, Monshizadeh M, Assi SM (2010) Semi automatic development of FarsNet; the Persian WordNet. In: Proceedings of 5th global WordNet conference, Mumbai, India
88. Kann V, Rosell M (2005) Free construction of a free Swedish dictionary of synonyms. In: Proceedings of NODALIDA 2005, Citeseer
89. Karthikeyan A (2010) Hindi English WordNet linkage. CSE Department, IIT Bombay, Bombay
90. Borin L, Forsberg M, Lönngren L (2013) SALDO: a touch of yin to WordNet's yang. Lang Resour Eval 47(4):1191–1211
91. Maamouri M, Graff D, Bouziri B, Krouna S, Bies A, Kulick S (2010) Standard Arabic morphological analyzer (SAMA) version 3.1. Linguistic Data Consortium, Catalog No.: LDC2010L01
92. Abdul-Mageed M, Diab MT (2011) Subjectivity and sentiment annotation of modern standard arabic newswire. In: Proceedings of the 5th linguistic annotation workshop. Association for Computational Linguistics
93. Buckwalter T (2004) Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2004L02. ISBN 1-58563-324-0
94. Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: International conference on intelligent text processing and computational linguistics. Springer

95. Balahur A, Turchi M (2014) Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput Speech Lang 28(1):56–75
96. Elkateb S, Black W, Rodríguez H, Alkhalifa M, Vossen P, Pease A, Fellbaum C (2006) Building a WordNet for arabic. In: Proceedings of the fifth international conference on language resources and evaluation (LREC 2006)
97. Turney PD (2001) Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: European conference on machine learning. Springer
98. Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R (1988) Using latent semantic analysis to improve access to textual information. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM
99. Stubbs M (2001) Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. Handb Discourse Anal 18:304
100. Kumar P, Jaiswal UC (2016) A comparative study on sentiment analysis and opinion mining. Int J Eng Technol 8(2):938–943
101. Passaro LC, Pollacci L, Lenci A (2015) Item: a vector space model to bootstrap an italian emotive lexicon. CLiC It 60(15):215
102. Kaity M, Balakrishnan V (2019) An automatic non-English sentiment lexicon builder using unannotated corpus. J Supercomput 1–26
103. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 417–424. https://doi.org/10.3115/1073083.1073153
104. Turney PD, Littman ML (2002) Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv:cs/0212012
105. Pozzi FA, Fersini E, Messina E, Liu B (2017) Chapter 1—Challenges of sentiment analysis in social networks: an overview. In: Sentiment analysis in social networks. Morgan Kaufmann, Boston, pp 1–11
106. Lafourcade M, Le Brun N, Joubert A (2016) Mixing crowdsourcing and graph propagation to build a sentiment lexicon: feelings are contagious. In: Metais E et al (eds) Natural language processing and information systems, NLDB 2016. Springer, Cham, pp 258–266
107. Yuang CT, Banchs RE, Siong CE (2012) An empirical evaluation of stop word removal in statistical machine translation. In: Proceedings of the joint workshop on exploiting synergies between information retrieval and machine translation (ESIRMT) and hybrid approaches to machine translation (HyTra). Association for Computational Linguistics
108. Al-Kabi MN, Kazakzeh SA, Ata BMA, Al-Rababah SA, Alsmadi IM (2015) A novel root based Arabic stemmer. J King Saud Univ Comput Inf Sci 27(2):94–103
109. Zhang Y, Tsai FS (2009) Chinese novelty mining. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 3. Association for Computational Linguistics
110. Abdul-Mageed M (2017) Modeling Arabic subjectivity and sentiment in lexical space. Inf Process Manag 56(2):291–307
111. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307
112. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations
113. Honnibal M, Montani I (2017) Spacy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (**to appear**)
114. Qiu X, Zhang Q, Huang X (2013) Fudannlp: a toolkit for chinese natural language processing. In: Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations, pp 49–54
115. Abdelali A, Darwish K, Durrani N, Mubarak H (2016) Farasa: a fast and furious segmenter for Arabic. In: HLT-NAACL Demos
116. Zhang H-P, Yu H-K, Xiong D-Y, Liu Q (2003) HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the second SIGHAN workshop on Chinese language processing-volume 17. Association for Computational Linguistics
117. Hussein DME-DM (2016) A survey on sentiment analysis challenges. J King Saud Univ Eng Sci
118. Bravo-Marquez F, Frank E, Pfahringer B (2016) Building a Twitter opinion lexicon from automatically-annotated tweets. Knowl Based Syst 108:65–78
119. Yue L, Chen W, Li X, Zuo W, Yin M (2018) A survey of sentiment analysis in social media. Knowl Inf Syst 1–47

120. Tang D, Wei F, Qin B, Zhou M, Liu T (2014) Building large-scale Twitter-specific sentiment lexicon: a representation learning approach. In: Proceedings of coling 2014, the 25th international conference on computational linguistics: technical papers, pp 172–182
121. Wang L, Xia R (2017) Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In: Proceedings of the 2017 conference on empirical methods in natural language processing
122. Kong L, Li C, Ge J, Yang Y, Zhang F, Luo B (2018) Construction of microblog-specific chinese sentiment lexicon based on representation learning. In: Pacific Rim international conference on artificial intelligence. Springer
123. Amir S, Astudillo R, Ling W, Martins B, Silva MJ, Trancoso I (2015) Inesc-id: a regression model for large scale twitter sentiment lexicon induction. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)
124. Dong X, de Melo G (2018) Cross-lingual propagation for deep sentiment analysis. In: Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI 2018). AAAI Press
125. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
126. Tang D, Qin B, Liu T (2015) Deep learning for sentiment analysis: successful approaches and future challenges. Wiley Interdiscip Rev Data Min Knowl Discov 5(6):292–303
127. Wang K, Xia R (2016) A survey on automatical construction methods of sentiment lexicons. Acta Automatica Sinica 42(4):495–511

Mohammed Kaity received the B.Sc. in Information Technology with an excellent grade, from Alandalus University for Science and Technology, Yemen, in 2008. He received M.Sc. in computer information system in 2012 from the Arab Academy for Banking and Financial Sciences. Currently, he is working toward the doctoral degree at the University of Malaya, Malaysia. He has worked in the academic field for more than 7 years. His research interests mainly include text classification, sentiment analysis, and social emotion classification.

Vimala Balakrishnan is an Associate Professor and a Fulbright Research Scholar affiliated with the Faculty of Computer Science and Information Technology, University of Malaya since 2010. She obtained her Ph.D. in the field of Ergonomics from Multimedia University, whereas her Master and Bachelor degrees were from University of Science Malaysia. She is also the current Program Coordinator for the Master of Data Science program offered at the Faculty of Computer Science and Information Technology. Dr. Balakrishnan's main research interests are in data analytics and sentiment analysis, particularly related to social media. Her research domains include healthcare, education and social issues such as cyberbullying. She has published approximately 55 articles in top indexed journals, 44 conference proceedings and six book chapters, has four patents and eight copyrights.