Artificial Intelligence Review (2019) 52:1415–1462
https://doi.org/10.1007/s10462-018-9670-y

A journey of Indian languages over sentiment analysis: a systematic review

Sujata Rani[1] · **Parteek Kumar**[1]

Published online: 12 December 2018

Abstract
In recent years, due to the availability of voluminous data on web for Indian languages, it has become an important task to analyze this data to retrieve useful information. Because of the growth of Indian language content, it is beneficial to utilize this explosion of data for the purpose of sentiment analysis. This research depicts a systematic review in the field of sentiment analysis in general and Indian languages specifically. The current status of Indian languages in sentiment analysis is classified according to the Indian language families. The periodical evolution of Indian languages in the field of sentiment analysis, sources of selected publications on the basis of their relevance are also described. Further, taxonomy of Indian languages in sentiment analysis based on techniques, domains, sentiment levels and classes has been presented. This research work will assist researchers in finding the available resources such as annotated datasets, pre-processing linguistic and lexical resources in Indian languages for sentiment analysis and will also support in selecting the most suitable sentiment analysis technique in a specific domain along with relevant future research directions. In case of resource-poor Indian languages with morphological variations, one encounters problems of performing sentiment analysis due to unavailability of annotated resources, linguistic and lexical tools. Therefore, to provide efficient performance using existing sentiment analysis techniques, the aforementioned issues should be addressed effectively.

Keywords Sentiment analysis · Opinion mining · Machine learning · Lexicon based · Indian languages · Aspect-based · Sentence-level · Systematic review

1 Introduction to sentiment analysis

In today's life, people always seek for suggestions and opinions of other people for their survival and decisions making. These opinions (or sentiments) benefit marketing and business operations in making products or services better. Recently there is a proliferation of World

✉ Sujata Rani
sujata.singla@thapar.edu

Parteek Kumar
parteek.bhatia@thapar.edu

[1] CSED, Thapar Institute of Engineering and Technology, Patiala, India

Wide Web sites that emphasizes user-generated content as users are the potential content contributors. Also, there are a lot of comments and blog-posts about trending activity on social media. The rapid evolution of microblogging and emerging sites such as Twitter has propelled online communities to flourish by enabling people to create, share and disseminate free-flowing messages and information (Asghar et al. 2018a). This explosion of opinionated text has fashioned an exciting area in text analysis, which is stated by many names like opinion mining, sentiment analysis, appraisal extraction, and subjectivity analysis (Pang et al. 2008). The terms sentiment analysis/opinion mining have been referred throughout this study. Sentiment Analysis (SA) is a natural language processing task to determine the user's attitude toward a particular entity by identifying and classifying users' opinions from a piece of text into different sentiments classes or emotions such as happy, sad, angry, or disgusted (Rani and Kumar 2017). The main aim of SA is to identify whether a text is objective or subjective. Objectivity specifies that the text does not consist of any opinion whereas subjectivity states that the text bears opinion content. For example, the sentence "This movie stars Amir Khan and Kajol." represents the objectivity as this sentence is a fact and conveys general information rather than an opinion or a view of some individual. And the sentence "This movie by Amir Khan and Kajol is superb." represents the subjectivity as this sentence represents opinion about the movie and the feelings of writer. The subjective text can be further categorized into the broad categories on the basis of the sentiments expressed in the text. Consider the sentence "I love to watch Star TV series." connotes the positive sentiment of writer about "star TV series" due to the sentiment word "love" and the sentence "The movie was awful." connotes the negative sentiment about movie because of sentiment word "awful". In the same way, the sentence "I usually get hungry by noon." is subjective sentence and connotes neutral sentiment as this sentence consists of feelings and views of user however, it does not consist of any positive or negative polarity.

1.1 Need of SA for Indian language text

India is a multi-lingual country with 22 official languages. Due to linguistic and cultural diversities of India, it has always left open a wide area for Natural Language Processing (NLP) researchers. The user base of Indian language users has already grown from 42 million in 2011 to 234 million in 2016 and it is expected to grow at a rate of 18% (to 536 million) by 2021 according to the KPMG-Google study (Onl 2017). The introduction of Unicode (UTF-8) has lead tremendous increase of web content. Mostly Indian people share their opinions on everything including products, latest movies, politicians, events, government rules, new trends, etc. in their own language as almost all the products, movies, newspapers and government official websites are now available in Indian languages. Presently, search engines such as Google also support Indian languages including Hindi, Bengali, Punjabi, Tamil, Telugu and Guajarati etc. The social media websites like Facebook and the micro blogging sites such as Twitter are also available in most of Indian languages. Thus, the escalation in the availability of voluminous Indian languages data has encouraged researchers to explore the research in this area.

1.2 Motivation for research

Some of the important points which motivated us to conduct this review are discussed as follows.

– India is a land of hundreds of languages. In recent times, Indian language data on web is increasing at an exponential rate. The SA of resource-poor Indian language data requires different lexical resources for different languages. This study presents various annotated datasets, pre-processing linguistic as well as lexical resources and frequently used SA techniques for different Indian languages.
– We realized the need of a systematic literature survey after analyzing the progressive research in area of SA for Indian languages. Therefore, the available research on the basis of a comprehensive and methodical search has been summarized in this study.

1.3 Our contributions

Our contribution for conducting this survey is summarized as follows.

– A systematic review technique has been followed to include the relevant research studies. The number of publications for all Indian languages has been categorized year-wise as well as on the basis of sources of extraction such as conferences, journals and workshops.
– Indian languages analyzed have been categorized on the basis of evolution of their research work for SA and percentage of research work carried out in a particular language.
– The online available annotated datasets, linguistic and lexical resources for different Indian languages are discussed to further carry out the research in these languages.
– A detailed analysis has been carried out to study various existing SA techniques for Indian languages accomplished by using available/developed resources.
– The amount of research work done by researchers at different sentiment levels, classes and domains to perform SA is also presented.
– Future research directions in the area of SA for Indian languages are also discussed in the last section.

1.4 Related surveys

Earlier surveys (Kaur and Saini 2014; Pandey and Govilkar 2015; Govindan and Haroon 2016) have been innovative, but these surveys cover only a limited set of languages and also don't follow any guidelines to conduct a systematic survey. As Kaur and Saini (2014) have surveyed on eight Indian languages. The research has been persistently growing in the field of SA for Indian languages, there is a need to carry out a systematic review to assess, upgrade, and assimilate the state-of-the-art research presented in this area. This study depicts a fresh and systematic literature survey covering all the Indian languages by keeping in mind the guidelines of a systematic review. Out of all Indian languages, the significant work reported on 13 Indian languages along with description of online available datasets for SA, pre-processing linguistic resources, approaches to create sentiment lexicons and sentiment classification techniques is covered in this study. This survey evaluates and discovers the research challenges on the basis of available existing research in the field of SA.

1.5 Article organization

The structure of the paper is structured as follows. Section 2 discusses about the history of sentiment analysis along with its general process, levels and applications. The review methodology followed to carry out this survey is covered in Sect. 3. Section 4 gives the brief description about the Indian language families and discusses about the differences between

Fig. 1 Sentiment analysis process

English and Indian languages for SA. Section 5 shows the extraction outcomes in the form of resources of publications, language-wise percentage of work and citations of research studies considered. Section 6 deals with the preliminaries for SA of Indian languages which consist of available annotated datasets, pre-processing linguistic resources and sentiment lexicons for Indian languages. Section 7 discusses about different SA techniques and evaluation measures adopted for validation. The status of work on SA for Indian language is presented in Sect. 8. Section 9 covers the findings identified from the survey and Sect. 10 concludes the paper and presents future indirections.

2 Evolution of sentiment analysis

The research works on sentiment analysis appeared as early as in 2000 (Pang et al. 2008). With advent of social media on internet, forum discussions, reviews, and its rapid growth, a huge amount of digital data (mostly opinionated texts, e.g., statuses, comments, arguments, etc.) were introduced, and to deal with this huge data, the SA field enjoyed a similar growth. Since early 2000, SA has become one of the most active research areas in NLP. However, most of the works are highly concentrated on the English language, favored by the presence of standard data sets.

2.1 Process of SA

Generally, SA process is performed through five phases as shown in Fig. 1. In the first phase, data are collected by using API's of social media sites or websites related to any particular domain. In data preparation phase, unstructured data are converted into a structured form by performing transliteration or by removing irrelevant and noisy content which is not useful for identifying sentiment.

In sentiment detection phase, computational tasks are performed to identify and extract the sentiment or opinion from the textual dataset. The fourth phase, i.e., sentiment classification classifies each subjective sentence into classification groups by using lexicon based, machine learning, deep learning or hybrid techniques etc. In this phase, the classifications groups identified can be further classified into different moods like gladness, happiness, pleasure, sorrow, regret, sadness etc. In presentation of output phase, the results of text analysis can be displayed by using graphical displays such as pie charts, bar charts, line graphs etc.

2.2 Levels of SA

Researchers have performed SA at three levels such as aspect/feature, sentence and document level as shown in Fig. 2.

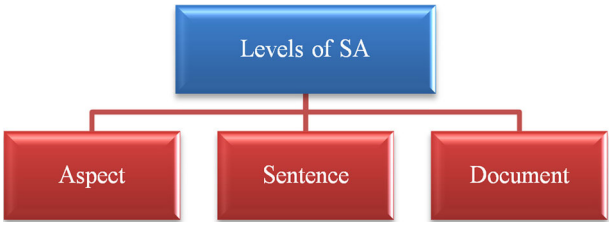

Fig. 2 Levels of SA

The brief description about these levels is given as follows.

(i) **Aspect level**: Aspect level is also known as Feature level. It performs finer-grained SA analysis. Aspect level targets at the opinion itself instead of taking language constructs into account. The basic aim of aspect-based SA is to detect aspects and identify the sentiments related to each aspect that are expressed by users (Asghar et al. 2017). It works on the idea that opinion consists of a target (of opinion) and sentiment about the target. For example, the sentence "Although the service is not that great, I still love this restaurant." clearly connotes the positive sentiment, however, this sentence is not entirely positive. In fact, this sentence represents the negative sentiment about the service, but positive sentiment about the restaurant.

(ii) **Sentence level**: At sentence level, SA is performed in two phases. In the first phase, the sentence is identified whether it is subjective (opinionated) or objective. Then opinionated sentence is further classified into different classes such as positive, negative or neutral in second phase. There may be different types of sentences such as direct, compound, comparative and sarcastic sentences for which performing SA becomes a difficult task. For example, the direct sentences represent the opinion about the object directly like "He is a good player." representing positive sentiment. Compound sentences are the sentences in which any compound (e.g., but) is used and the same sentence represents two different opinions. For example, "The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera" conveys both positive and negative opinions. And comparative sentences mean in which the similar objects are compared with each other. For example, "Ram's administration can't be compared with Ravan's administration." The sarcastic sentences have polarity opposite to their neighboring sentences.

(iii) **Document level**: At document level, it is assumed that the document consists of opinions about single entity only. Therefore, the task at this level is to classify whether it connotes a positive, negative or neutral sentiment. For example, given a latest movie review, SA system helps in determining whether the given review expresses the positive, negative or neutral sentiment about that movie.

2.3 Applications of SA

One of the important applications of SA is that it helps political parties or government to get an idea of what are the chances of their winning in coming elections and how much the public is satisfied with their policies. It can also help in predicting the success of a newly introduced social movement, for example, 'Odd-even car rule' in Delhi, 'Swachh Bharat Abhiyaan' etc. Public opinions can be gathered to have an idea of how much they are happy with the new change and based on that it can be determined whether the new policy is going to be successful

in future or not. In this way money, time and efforts can be saved or some necessary steps can be taken to make it more successful as people mostly express their opinions on social media platforms and the number of opinions grows during any change in policy or rule by government or any organization. One can predict the movie revenues in advance based on the opinions of other people. It can also help in education field to improve teaching quality and student learning which in turn can help the University administrators to take corrective measures (Rani and Kumar 2017).

Other important applications of SA include that it can be useful to product companies in increasing their revenues and customer retention by analyzing the opinions of people about their products or services. It is also very helpful in predicting customer trends and helps in developing more appealing and powerful marketing strategies. SA is a widely used in stock market predictions. Rises and falls in stock prices of a company are highly correlated with the sentiments expressed about that company on social media. Thus, based on the opinions expressed on social media one can decide whether in future the company is going to be in gain or loss and whether it is fruitful to invest money in stocks of that company (Arora 2013; Pang et al. 2008). The list of applications of SA is endless. A number of domains have been explored using sentiment analysis and still a lot more can be done using it.

3 Review methodology followed

The steps followed to conduct this review on SA for Indian languages are given as follows.

3.1 Development of review protocol

This systematic review has been conducted by identifying the related research studies from the renowned electronic databases as well as the topmost conferences related to the area. After this, to narrow down the count of selected studies, inclusion and exclusion criteria have been followed. Then, final research studies have been selected based on formulation of research questions and results have been compiled after performing in-depth analysis.

3.2 Research questions

The systematic review presented in this paper focuses on identifying and analyzing the existing literature survey describing different SA methods and techniques used for different Indian languages. It also finds the different lexical and lexicon resources as well as tools which are used by researchers to perform SA for Indian languages.

A set of research questions (listed in Table 1) have been formulated in order to conduct the systematic review in an efficient way.

3.3 Sources of information

A proper set of e-databases were chosen before starting the search process to identify the relevant research articles only. The electronic databases that were selected for identifying the research studies are Google Scholar (www.scholar.google.co.in/), Science Direct (www.sciencedirect.com), ACM Digital Library (www.acm.org/dl) and IEEE Xplore (www.ieeexplore.ieee.org). The most of the papers were published in topmost conferences related

Table 1 Research questions for systematic literature review

Research question	Motivation
RQ1: What is the year-wise status and which are the databases of publications since the inception of SA for Indian languages?	Identify the time frame and sources of publications in which the relevant large number of research studies have been published
RQ2: What is the impact of research studies considered?	Identify the research studies using citation information to include only the relevant work in this area
RQ3: Which Indian languages have been mostly explored till now?	Explore the Indian language families and identify languages for which majority of the SA research work has been carried out
RQ4: Which sentiment analysis techniques have been used mostly?	Identify the most commonly used SA techniques for each Indian language
RQ5: Which annotated datasets, linguistic and lexical resources are available online for Indian languages and which domains have been considered?	The online availability of annotated datasets, linguistic and lexical resources for SA are suggestive of ease of use and development of resources for other Indian languages. Also, identify the domains such as products, movies or social media platforms etc. in which corpora are available for Indian languages to perform SA
RQ6: What are the different factors considered while performing SA?	Identify the factors such as sentiment levels (such as aspect, sentence or document) and classes like positive, negative or neutral for which SA is performed
RQ7: Whether any online tools are available for SA of Indian languages?	Explore the availability of different SA tools that perform online SA for Indian languages
RQ8: What is the future indirections identified from the literature review?	Identify the unexplored relevant research visions

to NLP and linguistics; and also Google Scholar covers almost all the papers. The papers that were redundant on Science Direct, ACM Digital Library and IEEE Xplore have been excluded before final selection of research articles.

3.4 Inclusion and exclusion criteria

This systematic survey has been conducted by following the guidelines given by Kitchenham and Charters (2007). A systematical keyword-based advanced search has been followed to retrieve the significant research studies from the e-databases as shown in Table 2.

This systematical literature review consists of both qualitative and quantative research studies from 2010–2017 to ensure the completeness of review as attempt to work on SA for Indian languages was first commenced in 2010. The keywords "sentiment analysis" and "opinion mining" directed to a large number of results as this field is explored for different languages in different domains. The search has been performed in abstract and title using the search string "Sentiment analysis [in, of, for] [language_name]." For example, some research studies have considered the substring "in Hindi" or "for Hindi" or "of Hindi" in their title. Therefore, search has been performed by taking this into account so that all of the research studies in this field can be included. The research studies from various conferences, journals,

Table 2 Keyword-based advanced search

Source	Keyword	Publication type	Number of results	Time frame
Google Scholar	Sentiment analysis [in/of/for] [Language_name]	C/J/M	200	2010–2017
	Opinion mining [in/of/for] [Language_name]	C/J/M	43	
IEEE Xplore	Abstract: sentiment analysis [in/of/for] [Langauge_name]	C	26	
	Abstract: opinion mining [in/of/for] [Language_name]	C	10	
Science Direct	Abstract, Title, Keywords: sentiment analysis [in/of/for] [Language_name]	J	4	
	Abstract, Title, Keywords: opinion mining [in/of/for] [Language_name]	J	2	
ACM Digital Library	Abstract: sentiment analysis [in/of/for] [Language_name]	C/J	4	
	Abstract: opinion mining [in/of/for] [Language_name]	C	1	

Language_name: [Bengali, Hindi, Kannada, Konkani, Malayalam, Manipuri, Nepali, Oriya, Punjabi, Tamil, Telugu, Urdu], C: Conferences, J: Journals, M: Magazines

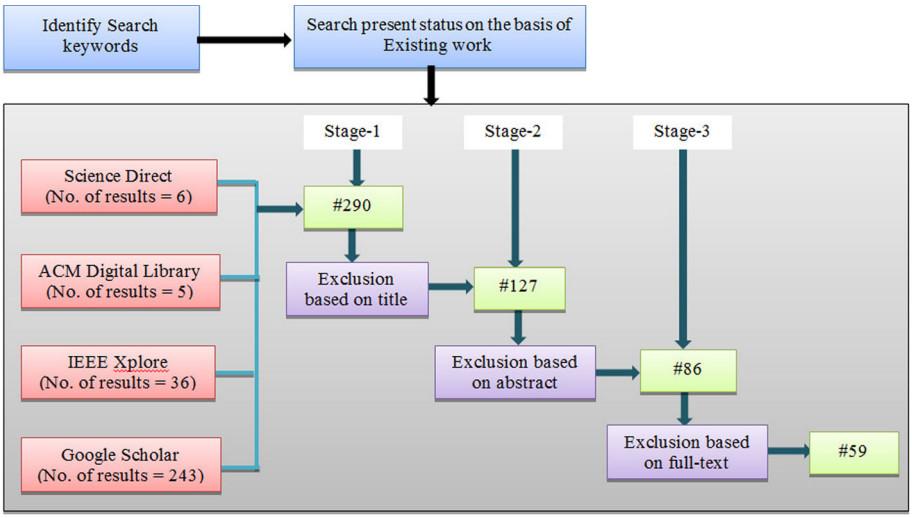

Fig. 3 Review technique followed

workshops along with masters and PhD thesis have been included by following an exclusion criterion at different stages shown in Fig. 3. Also, some individual searches have been applied on some conferences and journals related to NLP and linguistics to complete the e-search. Our search returned 290 research studies (shown in Fig. 3) which were reduced to 127 based on their titles, 86 based on their abstract and 59 on the basis of full-text. After that, these 59 research articles were analyzed in-depth to select a final list of research studies.

4 SA for Indian languages: the background

4.1 Introduction to Indian language families

Indian languages belong to several language families and broadly divided into four language families, i.e., Indo Aryan family (Arya), Dravidian family (Dravida), Sino Tibetan family (kirata) and Austroasiatic family (Nishada) as shown in Fig. 4.

Indo-Aryan language family covers about 74% of the Indian population and 24% of the total Indian population is covered by Dravidian languages. Austroasiatic and Sino-Tibetan languages are the language families' together covering 2% of the population (Ind 2015). The brief description about these language families is given as follows.

(i) **Indo-Aryan language family**: Indo-Aryan language family is part of the Indo-European family of languages and the mostly spoken language family in India. The utmost widely spoken languages of this language family are Hindi, Bengali, Punjabi, Odia (Oriya), Nepali, Konkani, Marathi, Gujarati, Sindhi, Assamese, Dogri, Urdu, Kashmiri and Sanskrit (Ind 2014).

(ii) **Dravidian language family**: It is the second largest language family in Indian language families. This language family is older than Indo-Aryan language family. The languages of this family are spoken mainly in southern and parts of eastern and central India as

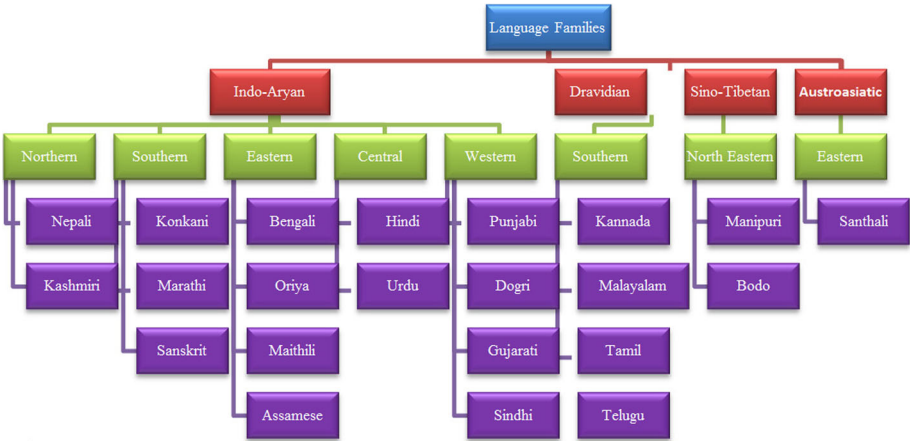

Fig. 4 Classification of Indian languages

well as in parts of north eastern Sri Lanka, Nepal, Pakistan and Bangladesh (Chand 2016). The major Dravidian languages are Telugu, Tamil, Kannada and Malayalam.

(iii) **Sino-Tibetan language family**: The Sino-Tibetan languages are referred as Kiratas in the oldest Sanskrit literature. This language family is also older than the Indo-Aryan language family. These languages have three major sub-divisions such as The Tibeto Himalayan, The North Assamese and The Assam–Myanomari (Burmese). The main Sino-Tibetan languages are Manipuri and Bodo.

(iv) **Austroasiatic language family**: The Austric languages are referred as Nisadas in the oldest Sanskrit literature of India and these languages are mainly spoken in the central, eastern and north-eastern India. This ancient language family came into existence before the arrival of Aryans. The most spoken language of this family is Santhali.

Mainly the research work in the field of SA has been done in English as well few other non-English languages such as Arabic, Chinese, etc. A very less contribution exists for Indian Languages such as Hindi, Tamil, Telugu, Bengali, etc. (Kaur and Saini 2014). The primary reason behind this is the lack of annotated datasets, linguistic as well as lexical resources and tools for Indian languages.

4.2 Evolution of Indian languages for SA

The evolution of Indian languages in the field of sentiment analysis started in 2010 when Joshi et al. (2010) first attempted to work on SA for Indian languages. The authors performed sentiment analysis for Hindi language and later on researchers started working on different Indian languages such as Bengali, Tamil, Telugu, etc. The year-wise evolution of Indian languages for SA is shown in Fig. 5.

4.3 Differences between sentiment analysis in English and Indian languages

It is well-known that different languages have their own unique ways of expression. The basic difference between English and Indian languages is the language structure. For example, English has an SVO (Subject Verb Object) structure, while Hindi follows an SOV (Subject

Fig. 5 Evolution of Indian languages for SA

Object Verb) structure. This basic structural difference between English and Indian languages has consequences in deciding the polarity of a text. The same set of words with slight variations and changes in the word order affect the polarity of the words in the text. Therefore, a deeper linguistic analysis is required while dealing with the Indian languages to perform SA. For example, consider the sentences given below which represent the difference between language structure of English and Hindi.

English: Peter is playing cricket.
S V O
Hindi: पीटर क्रिकेट खेल रहा है।
S O V

The above sentences clearly indicate that English sentences follows the SVO word order only, while the Indian language sentences don't follow any word order. The freely word order nature of sentences of Indian language makes the pre-processing difficult. Despite of language structure, there are some other differences between English and Hindi language which make the SA process difficult. These differences (Arora 2013) are discussed as follows.

(i) **Null-subject divergence**: A null subject language in linguistic topology is a language in which grammar permits an independent clause known as "null subject" to lack an explicit subject. Some of the null subject Indian languages are Hindi, Tamil, and Telugu etc. whereas English obligatorily requires a subject. Due to this null-subject divergence, SA process becomes difficult. For example, consider the following sentence which represents the null subject divergence between English and Hindi.

English: Long ago, there was a king.
Hindi: बहुत पहले एक राजा था।
Long ago one king was

(ii) **Handling spelling variations**: In case of Indian languages, the same word with same meaning can occur with different spellings, so it's quite complex to have all the occurrences of such words in a lexicon and even while training a model it's quite complex to handle all the spelling variations. For example, consider the following sentence which shows that the word 'costly' can be written in Hindi with different spelling variations.

English: This phone is very costly.
Hindi: यह फोन बहुत मंहगा/महंगा/महँगा है।

(iii) **Morphological variations**: Handling the morphological variations is also a big challenge for Indian languages. Indian languages are morphologically rich which means that lots of information is fused in the words as compared to the English language

where another word is added for the extra information. For example, in the following sentence, the verb 'kill' carries far much information apart from just the root. It carries the inflection which provides information/idea about the tense, gender and person. Thus, with same root there can be many words in a language with varying information i.e., multiple variations of same words can have the same root with respect to the sense of tense, gender, person and other information.

English: Ram killed Ravana. Hindi: राम ने रावण को मारा ।
English: Ram is killing Ravana. Hindi: राम रावण को मार रहा है।

(iv) **Paired words**: Sometimes paired words are used in Indian language context. These paired words can be combination of two different opposite, meaningful and meaningless words. For example, the word 'tease in the following sentence is specified by combining two different words in Hindi.

English: Karan teased Sangeeta. Hindi: करण ने संगीता के साथ छेड-छाड की।

(v) **POS divergence**: As in case of SA, mostly adjectives consists of sentiment in a text. However, sometimes the POS of a word get changed while performing its translation from English to target language. Consider the following sentences in which the words acting as adjectives in English become adverbs or verbs after translation into Hindi language.

English: The children watched in wide-eyed amazement. (Adjective) Hindi: बच्चे आश्चर्य से आंख फाडे देख रहे थे। (Adverb)
English: He was in a bad mood at breakfast and wasn't very communicative. (Adjective) Hindi: नाश्ते के समय वह खराब मूड में था और ज्यादा बात चीत नहीं कर रहा था। (Verb)

These different structural and grammatical challenges of Indian languages make the SA task harder. To understand these rich variations of attributes of the Indian context words, the system needs robust morph analyzer so that the right sense of the word can be mined. Notably, efficient linguistic resources are required to pre-process Indian language context and to take care of spelling and multilingual issues.

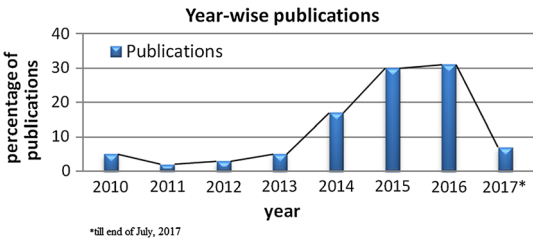

Fig. 6 Year-wise publications on SA for different Indian languages

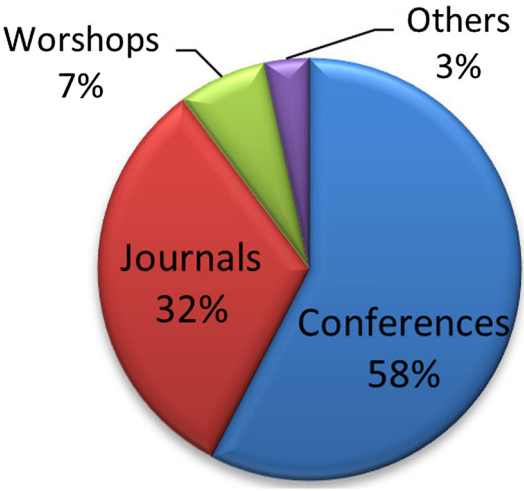

Fig. 7 Status of publications from different sources

5 Extraction outcomes

The aim of this work is to identify the available research on SA for Indian languages and is stated in Table 1 in the form of research questions. To answer the research question **RQ1**, year-wise status and origin of sources of publications on SA for Indian languages have been explored which are represented in Figs. 6 and 7 respectively.

The first attempt in this area was commenced in 2010 therefore; the year-wise status of publications from 2010 to 2017 is depicted in Fig. 6. From the figure, it has been analyzed that research in this area is continuously growing from the last couple of years.

While in-depth analysis, it has also been observed that most of the research articles on Indian sentiment analysis are published in a extensive variety of conference proceedings and journals. The approximate 58% of the research articles are published in conferences, 32% in journals, 7% in workshops, and remaining 3% is covered by thesis and online reports as shown in Fig. 7. The highest percentage of research publications came from conferences, followed by journals.

To address the research question **RQ2**, data has been analyzed and is responded through the Table 3. From this table, it can be observed that approximate seventy five percent of the research studies considered in this survey has less than equal to five citations, i.e., only the relevant research studies have been considered. The highest cited research study is the work done by Joshi et al. (2010) which has more than fifty citations. The authors of this research study set the benchmark in this area and other researchers followed the approaches discussed by them to perform SA for other Indian languages.

Table 3 Number of publications according to citations

Citation count	≤ 5	> 5 and ≤ 20	> 20 and ≤ 50	> 50
No. of publications	44	10	4	1

6 Preliminaries for SA of Indian languages

6.1 Dataset

The first phase to perform SA is dataset collection. Mainly the social media platforms such as Twitter, blogs, discussion forums and review sites related to products, movies and travels have been used to perform SA for Indian languages. Some of the annotated datasets of tweets and reviews are available online for four Indian languages such as Hindi, Bengali, Tamil and Marathi as given in Table 4. These dataset are annotated into different classes such as positive, negative and neutral. As majority of research work on SA has been done for Hindi language therefore, aspect and sentence level annotated datasets are available for Hindi language across various domains. The information given in Table 4 helps in attaining the answer to the research question **RQ5**. This table provides the summary about online available annotated datasets for Indian languages.

6.2 Pre-processing linguistic resources

Some of the pre-processing resources such as shallow parser, Part Of Speech (POS) tagger, dependency parser and morphological analyzer to perform SA for Indian languages along with their online availability are given in Table 5. The brief description about these resources is given as follows.

(i) **Shallow parser**: Generally, shallow parser provides the analysis of a sentence in the form of morphological structure, Chunking, POS tagging, etc. Shallow parsers for Indian languages are developed under a consortium project funded by Government of India (sha 2012). Till now, these are mainly available online for nine Indian languages.
(ii) **Morphological analyzers**: Morphological analyzers give the root word and other features such as gender, number, tense etc. Thus, morphological analysis is the process of imparting grammatical information of a word given its suffix. The independent morphological analyzers are available online for five Indian languages. However, one can also use the shallow parser to perform morphological analysis.
(iii) **POS tagger**: POS tagging is a process of classifying and labeling the words of a sentence according to their POS information which includes nouns, verbs, adjectives, determiners, adverbs, and so on. POS tagger generally indicates the status of the word based on the morphological and/or syntactic properties of a language. POS taggers are independently available online for four Indian languages. However, one can also use the shallow parser to extract the POS tagging information.
(iv) **Dependency parser**: Dependency parsing is the process of revealing the dependency tree of a sentence through labeled links which represent the dependency relationships between words. Researchers have worked on creation of dependency parsers for various Indian languages such as Telugu, Tamil, Bengali, etc. but presently dependency parser is available online only for Hindi language.

Table 4 Summary about annotated datasets for SA of Indian languages

Sr. no.	Language (Author)	Dataset type	Level	Dataset details	Link
1	Hindi (Patra et al. 2015)	Tweets	S	168 positive, 559 negative, 494 neutral tweets	http://amitavadas.com/SAIL/index.html
2	Hindi (Akhtar et al. 2016b)	Movie and Product Reviews	A	2290 positive, 712 negative, 2226 neutral and 189 conflict reviews	http://amitavadas.com/SAIL/index.html
3	Hindi (Akhtar et al. 2016a)	Movie and Product Reviews	A	2250 positive, 635 negative, 2241 neutral and 128 conflict reviews	https://www.iitp.ac.in/~ai-nlp-ml/resources.html
4	Hindi (Akhtar et al. 2016c)	Movie Reviews	S	823 positive, 530 negative, 598 neutral, 201 conflict reviews	https://www.iitp.ac.in/~ai-nlp-ml/resources.html
5	Hindi (Bakliwal et al. 2012)	Product Reviews	S	350 positive, 350 negative	Available on request to author
6	Bengali (Patra et al. 2015)	Tweets	S	277 positive, 354 negative, 368 neutral tweets	http://amitavadas.com/SAIL/index.html
7	Tamil (Patra et al. 2015)	Tweets	S	387 positive, 316 negative, 400 neutral tweets	http://amitavadas.com/SAIL/index.html
8	Marathi (Balamurali et al. 2012)	Tourism	S	75 positive, 75 negative reviews	http://www.cfilt.iitb.ac.in/resources/senti/MPLC_tour_downloaderInfo.php
9	Hindi (Balamurali et al. 2012)	Travel	S	100 positive and 100 negative reviews	http://www.cfilt.iitb.ac.in/resources/senti/HPLC_tour_downloaderInfo.php

S: Sentence, A: Aspect

(v) **Sandhi splitter**: Sandhi-Splitter is a computational tool which shows all possible splitting of a given string. Currently, Sandhi splitter is available online for Malayalam language.

The research studies have been analyzed and on the basis of that, the answer to the research question **RQ5** has been addressed. Table 5 summarizes the details about the online available pre-processing linguistic resources for different Indian languages.

6.3 SentiWordNet (SWN): lexical resource for SA

Researchers have either manually constructed or used WordNets to create lexical resources for SA. SWN is such a lexical resource that is mostly used for SA. SWN is the result of annotation of all WordNet synsets on the basis of degrees of polarity, i.e., positivity, negativity, and neutrality (Baccianella et al. 2010). WordNets have also been created for a number of Indian languages. For example, Indo WordNet is a linked structure of WordNets of all major Indian languages and currently supports 19 Indian languages (Bhattacharyya 2017).

Generally, WordNet and bi-lingual dictionary based approaches are followed for creation of SWN(s). In WordNet based approach, SWN for the target language is developed by mapping the sysets of English SWN along with polarity scores into target language sysets using Indo WordNet. In bi-lingual dictionary based approach, the polarity scores are extracted from English SWN and assigned to the words of target language. Das and Bandyopadhyay (2010c) proposed three other approaches such as corpus based, antonym based and gaming technology to increase the coverage of developed SWN. Presently, SWN(s) are available online for three languages such as Hindi, Bengali and Telugu at http://www.amitavadas.com/sentiWordNet.php.

Table 6 provides the answer for the research question **RQ5** as it gives the information about the online available SWN(s) for different Indian languages along with their development approach and count of synsets.

7 SA techniques and evaluation measures used

7.1 SA techniques

From the comprehensive survey, it has been observed that SA techniques can be classified into three categories such as lexicon based, machine learning and deep learning as shown in Fig. 8. The brief description about these techniques is given as follows.

7.1.1 Lexicon based

This approach is also known as rule-based approach. In this approach, certain rules are followed along with the use of sentiment lexicons to determine the sentiment from the text. The sentiment lexicon consists of words along with their sentiment polarity, e.g., "excellent" as positive, "horrible" as negative (Rehman and Bajwa 2016; Syed et al. 2010). The sentiment orientation of an unknown document is computed by matching the words in the document to words in the sentiment lexicon, and then taking the aggregate of their values using one of various algorithms. The positive/negative values of the words in the text are aggregated which help in producing the semantic orientation for the entire text. Mainly the three approaches such as manual construction, dictionary-based and corpus-based are followed to construct

Table 5 Online available pre-processing linguistic resources

Sr. no.	Resource	Languages	Online link
1.	Shallow parser	Hindi, Urdu, Punjabi, Tamil, Bengali, Kannada, Telugu, Malayalam and Marathi	http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
2.	Independent morphological analyzers	Hindi, Marathi, Kannada, Punjabi and Telugu	http://ltrc.iiit.ac.in/showfile.php?filename=onlineServices/morph/index.htm, http://www.learnpunjabi.org/punjabi_mor_ana.asp
3.	Independent POS tagger	Hindi, Kannada, Punjabi and Telugu	http://sivareddy.in/downloads, http://punjabi.aglsoft.com/punjabi/?show=tagger
4.	Dependency parser	Hindi	https://bitbucket.org/sivareddyg/hindi-dependency-parser
5.	Sandhi splitter	Malayalam	https://github.com/libindic/sandhi-splitter

Table 6 Summary about SWN(s) created for Indian Languages

Language family	Language	Approach	Lexical resources	Synsets (approx.)	Ref
Indo-Aryan	Hindi	WordNet	Indo WordNet (Bhattacharyya 2017), English SWN (Esuli and Sebastiani (2007))	16,000	Bakliwal et al. (2012)
		Dictionary	SHABADKOSH, Shabdanjali	22,708	Das and Bandyopadhyay (2010c)
	Bengali	Dictionary	Samsad Bengali-English dictionary	34,117	Das and Bandyopadhyay (2010b)
	Konkani	WordNet	Indo WordNet (Bhattacharyya 2017), Hindi SWN	368	Fondekar et al. (2016)
	Nepali	Dictionary	English SWN (Esuli and Sebastiani 2007), English-Nepali Dictionary	629,930	Gupta and Bal (2015)
	Punjabi	Dictionary	Hindi SWN	7860	Kaur and Gupta (2014b)
Dravidian	Telugu	Dictionary	Charles Philip Brown English-Telugu Dictionary, Aksharamala English-Telugu Dictionary, English-Telugu Dictionary	30,889	Das and Bandyopadhyay (2010c)
	Kannada	Manually	Hindi SWN	5043	Deepamala and Kumar (2015)
	Malayalam	Dictionary	Hindi SWN	2000	Anagha et al. (2014)

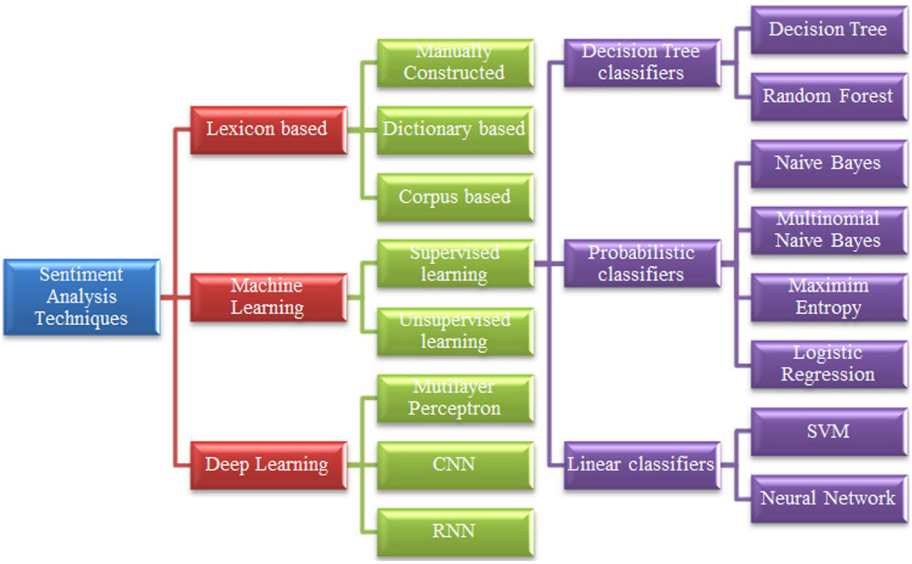

Fig. 8 SA techniques for Indian languages

the sentiment lexicon (Joshi et al. 2010). The manual construction approach is difficult and time consuming. In this approach, polarities are manually assigned to sentiment words by humans. Dictionary-based is an iterative approach in which small set of sentimental words are selected initially and this set then iteratively grows by adding the synonyms and antonyms from the WordNet. This iterative process continues till no new words are remaining to be added to the seed list. Corpus-based techniques depend on on syntactic patterns in large corpora and can produce sentiment words with relatively high accuracy.

7.1.2 Machine learning (ML)

Machine learning is a subfield of artificial intelligence that makes computers learn without being explicitly programmed. It is basically constructing the algorithms that can learn from data and make predictions on the related data. ML is categorized into two classes namely supervised and unsupervised machine learning. In supervised ML, there is a predetermined set of classes into which the documents are classified and training data is available for each class. The system uses any of the classification algorithms such as Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), k-Nearest Neighbor (k-NN) and trains a model from the given data. This trained model is then used for making predictions and assigning the documents into different sentiment classes (Se et al. 2016). In case of unsupervised approach of ML, no labeled data is provided to models. This approach works on the basis of computation of Semantic Orientation (SO) of specific phrases within the text. If the average SO of phrases is above some predefined threshold, the text is classified as positive and otherwise, it is specified as negative.

(i) **Naive Bayes** Naive Bayes classifier belongs to the family of probabilistic classifiers and is based on Bayes theorem. It takes the probability distribution of words in the training dataset and assumes them to be mutually independent. Given a feature vector $(x_1, x_2, x_3, \ldots, x_n)$ and a class variable y, Naive Bayes assigns the class to feature

vector according to the Bayes formula given in (1).

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(y)P(x_1, x_2, \ldots, x_n|y)}{P(x_1, x_2, \ldots, x_n)} \tag{1}$$

In this formula, $P(y|x_1, x_2, \ldots, x_n)$ represents the posterior probability. $P(y)$ is prior class probability and $P(x_1, x_2, \ldots, x_n)$ is the prior probability of feature set. These prior probabilities are obtained from training dataset. $P(x_1, x_2, \ldots, x_n|y)$ represents the conditional probability of feature vector $(x_1, x_2, \ldots, x_n)$ given the class y. The formula can be generalized as given the feature vector, Naive Bayes finds the probability of each class to be assigned to this feature vector and assigns the class with maximum probability. It assumes mutual independence among the features as shown in (2).

$$P(x_1, x_2, \ldots, x_n|y) = \prod_i P(x_i|y) \tag{2}$$

(ii) **Multinomial Naive Bayes (MNB)** Multinomial Naive Bayes is a specific version of Naive Bayes. Whereas a simple NB classifier models the document as the presence or absence of words, MNB takes into account the words counts. Given a class c, MNB estimates the conditional probability of a particular word as the relative frequency of the word in that class as given in (3).

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \epsilon V} T_{ct'}} \tag{3}$$

here t is the term/word and c is the class under consideration. This formula calculates the probability of a word to be classified into a class as the count of that word in the class with respect to count of all the words in that class.

(iii) **Maximum entropy (ME)** Maximum Entropy classifier is similar to NB classifier except that it doesn't make any assumption about the independence of features. The principle idea behind Maximum Entropy is that it tries to maximize the Entropy and at the same time satisfying the constraints specified. The idea behind Maximum Entropy is to have a model that is as unbiased as possible and thus the probability distribution to be as uniform as possible. Maximum Entropy is when all the events are equally likely to occur and have maximum uncertainty. The formula for Entropy is given in (4) and the goal is to maximize $H(P)$.

$$H(P) = \sum p(a, b)log(p(a, b)) \tag{4}$$

(iv) **Support vector machines** Support Vector Machines work on the concept of a decision plane or a hyperplane. It tries to find a hyperplane which separates the data belonging to two classes as far apart as possible as represented in (5).

$$(\overrightarrow{w} . \overrightarrow{x}) = \sum_i y_i \alpha_i (\overrightarrow{x_i} . \overrightarrow{x}) + b \tag{5}$$

here $\overrightarrow{x_i} = (x_{i1}, \ldots, x_{in})$ is input feature vector, y_i is output class, $\overrightarrow{w_i} = (w_{i1}, \ldots, w_{in})$ is the weight vector defining the hyperplane and α_i is Lagrangian multiplier. Once the hyperplane is constructed, the class of any feature vector can be determined. Figure 9 shows the working of Support Vector Machine.

(v) **Logistic regression (LR)** Logistic Regression is a multi-class logistic model which is used to estimate the probability of a response based predictor variables in which there are one or more independent variables that determine an outcome. The expected values

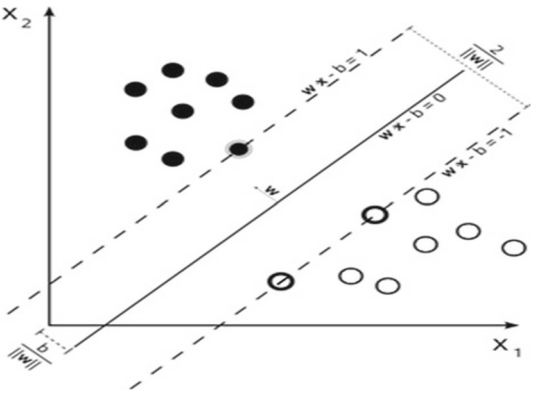

Fig. 9 Support vector machines

of the response based predictor variable are formed based on combination of values taken by the predictors.

(vi) **Decision tree (DT)** Decision Tree is a decision support tool that uses a treelike model for the decisions and likely outcomes. A decision tree is a tree in which each internal (non-leaf) node is labeled with an input feature and each leaf of the tree is labeled with a class.

(vii) **Random forest (RF)** Random Forest is an ensemble of Decision Trees. Random Forests construct multiple decision trees and take each of their scores into consideration for giving the final output. Decision Trees tend to overfit on a given data and hence they will give good results for training data but bad on testing data. Random Forests reduce overfitting as multiple decision trees are involved.

7.1.3 Deep learning

Deep learning is a branch of machine learning inspired from human brain. It has emerged as a powerful approach for pattern recognition and language processing in recent years. Because of its ability to automatic feature engineering and appreciable accuracy, it is getting widespread popularity these days. Deep learning refers to the number of layers that comprise the Neural Network (NN). Early NNs were defined with three layers; input, hidden, and output. Adding several hidden layers makes the NN 'deep' and enables it to learn more subtle and complex relationships. The number of 'hidden layers' decides how deep the network is. Neural networks cannot process direct words, but they work on word embeddings or more specifically feature vectors representing those words. One of the ability of deep learning is that for feature learning, handcrafted features are replaced with efficient algorithms (Seshadri et al. 2016). These are capable of capturing very high level features from input data. As neural networks learn features from the task in hand they can adapt to any domain. Deep nets can perform better than traditional machine learning approaches if the sufficient amount of training data is given (Akhtar et al. 2016c). Deep learning has found applications in a number of areas like sentiment analysis, computer vision, automatic speech recognition etc. The brief description about some of the important deep learning models is given as follows.

(i) **Multilayer perceptron model** A Multi-Layer Perceptron model is a feed forward supervised Artificial Network (ANN) model that learns a function given as follows by training on a dataset. $f(.) : R_m \rightarrow R_o$ Where m represents the number of input dimensions

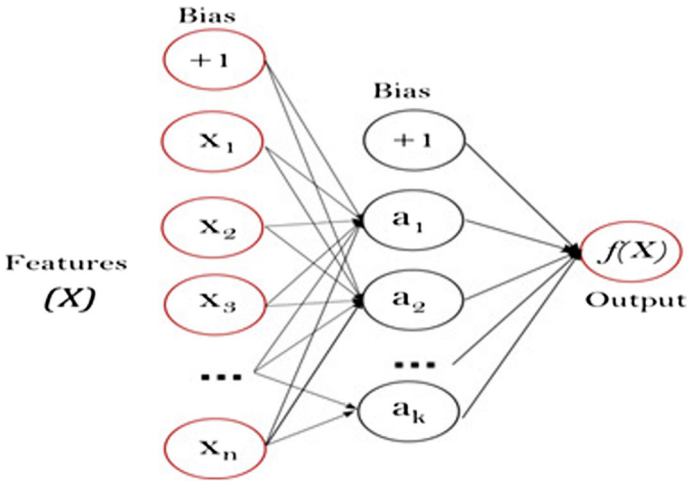

Fig. 10 Multi-layer perceptron model

and o is the number of output dimensions. For a classification problem, the number of nodes in input layer depends upon the length of input vector and number of nodes in output layer depends upon the number of pre-defined classes. There can be any number of hidden layers in between input and output layer. Input layer takes $(x_1, x_2, \ldots, x_m)$ as input vector. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $(w_1x_1 + w_2x_2 + ... + w_mx_m)$ followed by a non-linear activation function given as follows. $g(.) : R \rightarrow R$ The output layer then transforms values from the last hidden layer into output. Figure 10 shows the working of Multi-Layer Perceptron model.

(ii) **Convolutional neural network (CNN)** Convolutional Neural Networks (CNNs) are very much similar to the ordinary neural networks. Like ordinary neural networks, neurons in CNNs take some input, process it and propagate it further. The difference is that convolutional neural networks explicitly assume input as images. This is the reason they are explicitly used for analyzing image data. Regular neural networks don't scale well to full images. For small dimensions these are manageable, but as the dimensions grow, more neurons and parameters are required leading to the problem of over fitting. As CNN is specifically designed for image data it constrains the architecture in a more sensible manner. Unlike regular neural networks, neurons in each layer of CNN are arranged along three dimensions, i.e., height, width and depth. A CNN has three types of layers namely convolutional layer, pooling layer and fully-connected layer. Convolutional layer is the main building block of a CNN as most of the computations are done at this layer. Figure 11 displays the architecture of Convolutional Neural Network. The CNN architecture shown above consists of 4 layers. First is input layer that represents the sentences over $n * k$ dimension, second is convolutional layer, then max pooling layer and finally fully connected layer producing output results.

(iii) **Recurrent neural network (RNN)** In a regular neural network, all inputs and outputs in a layer are considered independent. This is the reason that from the present state future events cannot be predicted and this is the major shortcoming of an ordinary neural network. This problem is resolved by Recurrent Neural Networks. RNNs make use of loops making information to persist. Thinking another way, RNNs have memory which stores the information calculated so far and using that information for future predictions.

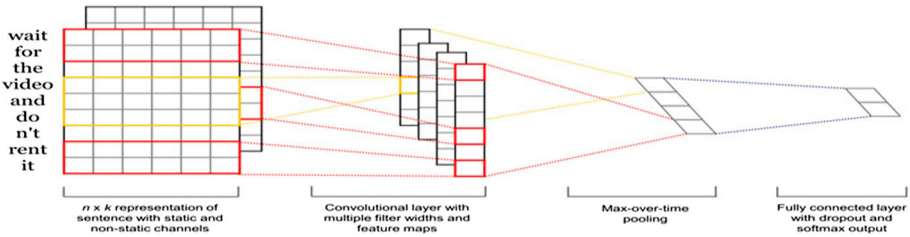

Fig. 11 Convolutional neural network

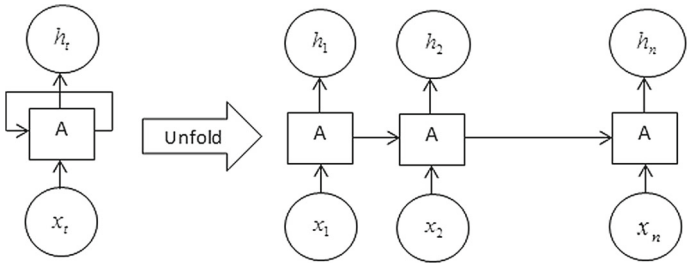

Fig. 12 Recurrent neural network

RNNs have applications in a number of areas which an ordinary neural network cannot solve, for example, based on the current events in a movie, RNN can determine the next event. Similarly, given a sequence of words, the next word in sequence can be determined using RNN. Other applications include handwriting recognition and speech recognition. The most common Recurrent Neural Network is Long Short Term Memory (LSTM) in short. The principle architecture of RNN is shown in Fig. 12. In the above figure, h_t is input and x_t is corresponding output of neural network. As RNN is unfolded, it becomes similar to regular neural network.

Out of the above discussed deep learning techniques, it has been observed that researchers have mostly used RNN as it gives better results as comparison to other techniques.

7.2 Evaluation measures

Mostly the researchers have used Accuracy (A) as evaluation measure. However, accuracy is not a sufficient metric to evaluate the efficiency and effectiveness of a classifier. Therefore, some of the researchers have also used the other metrics such as Precision (P), Recall (R) and F-measure (F) in addition to A as these metrics provide much greater insight into the performance features of a classifier. For a sentiment classifier, these four metrics can be defined in terms of True Positive (t_p), False Positive (f_p), True Negative (t_n) and False Negative (f_n) rates. Here, t_p rate represents the positive review and classifier also classifies it as positive, t_n rate represents the negative review and classifier also classifies it as negative. f_p represents the positive review but classifier classifies is as negative, f_n represents the negative review but classifier classifies it as positive. The brief description about the other evaluation measures is given as follows.

(i) **Accuracy** Accuracy A can be defined as in terms of how close the review classification suggested by the classifier is, to the actual sentiments present in the review.

(ii) **Precision** The precision P can be defined as in terms of the exactness of a classifier. A higher P means less false positive and vice versa. $P = \frac{t_p}{(t_p+f_p)}$

(iii) **Recall** The recall R can be defined as in terms of the sensitivity or completeness of the classifier. Higher R means less false negative and vice versa. $R = \frac{t_p}{(t_p+f_n)}$

(iv) **F-measure** F-measure is measured by combining Precision and Recall, which is the weighted harmonic mean of both values, defined as follows. $F = \frac{2PR}{(P+R)}$

8 Status of SA work for Indian languages

As far as development of SA systems with respect to Indian languages is concerned, most of the research work done in this domain is for English language and European languages. The research work reported in the field of SA for all Indian languages for different language families is presented in this section. The majority of the research work on SA has been performed for Indo-Aryan languages (such as Hindi, Bengali and Urdu) and Dravidian languages (such as Tamil, Malayalam and Kannada). The brief description about the research being performed on these languages is given as follows.

8.1 Languages with major research work

(i) **Hindi** Joshi et al. (2010) first attempted to work on SA for Indian languages. The authors proposed a fallback strategy which follows three approaches such as In-language SA, machine translation and resource based SA by developing own lexical resource Hindi SentiWordNet (HSWN). The authors achieved an accuracy of 78.14%. Balamurali et al. (2012) used WordNet sense-based features and experimented on a dataset travel reviews to perform SA. Bakliwal et al. (2012) developed Hindi subjective lexicon using bi-lingual dictionary and translation based approach. The authors performed sentiment classification using this lexicon and achieved an accuracy of 79%. Mittal et al. (2013) performed SA of movie reviews and also handled negation as well as discourse relations. The authors used HSWN to perform SA and achieved an accuracy of 80.21%. Arora (2013) performed the sentiment analysis on a corpus of Hindi reviews and blog related to products and movies using subjective lexicon and n-gram approaches. Bansal et al. (2013) used deep learning to perform SA of movie reviews and achieved an accuracy of 64%. Sharma et al. (2014) proposed a SA system using an unsupervised dictionary based approach and classified movie reviews into three categories, such as positive, negative and neutral. Their proposed methodology handles negations also and the system has achieved an accuracy of 65%. Sharma and Bhattacharyya (2014) proposed a bootstrap approach to extend the HSWN using existing HindiWordNet for all the four parts of speech, i.e., noun, adjective and verb. The authors used this lexicon to validate the SA system over movie and product reviews domain, and achieved an accuracy of 87%. Prasad et al. (2015) performed sentiment classification of tweets using decision tree under constrained and unconstrained environment. Venugopalan and Gupta (2015) used tweet specific features and performed SA using ML algorithms SVM and DT. Kumar et al. (2015c) performed SA of Hindi tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed sentiment classification using a regularized least square method. Kumar et al. (2015a) performed SA of Hindi tweets using SVM and MNB classifier. The authors also constructed their own lexicon, namely DT_COOC lexicon

using distributional thesaurus and sentence co-occurrences. Sarkar and Chakraborty (2015) also performed SA of Hindi tweets using SVM and MNB classifier. Se et al. (2015) reported the work on SA for Hindi tweets using machine learning classifiers and classified them into positive, negative and neutral class. The authors analyzed that machine learning classifiers perform better within constrained environment, i.e., without the availability of NLP tools like POS tagger, Named Entity Recognition (NER). Jha et al. (2015) proposed a Hindi opinion mining system and used NB classifier and unsupervised approach of POS tagging to perform SA. (Pandey and Govilkar 2015; Sharma et al. 2015) used unsupervised lexicon based approach to perform SA using HSWN and classified sentences into positive, negative and neutral class. The authors also handled negations and discourse relations. Seshadri et al. (2016) performed SA of Hindi tweets on Sentiment Analysis in Indian Languages (SAIL-2015) dataset. Patra et al. (2015) used RNN and classified tweets into positive, negative and neutral class. Phani et al. (2016) also used the SAIL-2015 dataset to perform SA for Hindi language. The authors experimented using six classifiers such as NB, LR, DT, RF, SVM using four categories of features, namely word n-grams, character n-grams, surface and SWN features. Sharma and Moh (2016) predicted Indian election results of 2016 using lexicon based and machine learning approaches by collecting tweets in Hindi language and analyzed that ML techniques perform better than lexicon based. Akhtar et al. (2016b) performed aspect-based SA in Hindi and developed an annotated dataset consisting of Hindi product reviews. The authors used SVM classifier and attained an accuracy of 54.05%. The authors also experimented in four domains 'electronics', 'mobile apps', 'travels' and 'movies' using three classifiers such as NB, DT and SMO in MEKA (a Multi-label/Multi-target Extension to WEKA) and reported that NB performs better in electronics and mobile apps domain, while decision tree reports better results for the travels and movies domain (Akhtar et al. 2016a). Akhtar et al. (2016c) first attempted to work on SA for Hindi using deep learning based model such as CNN and performed SA at both aspect and sentence level.

(ii) **Bengali** Das and Bandyopadhyay (2010a) experimented on Bangla news text to find the polarity of opinions using SVM classifier and classified opinionated phrase as either positive or negative and attained a precision of 70.04% and a recall of 63.02%. Hasan et al. (2014) performed SA on Bangla text using contextual valency analysis. The authors used SWN and WordNet to find the prior valence of Bangla words. Kumar et al. (2015c) performed SA of Bengali tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed SA using a regularized least square method. Kumar et al. (2015a) performed SA of Bengali tweets using SVM and MNB classifier. The authors also constructed their own lexicon, namely DT_COOC lexicon using distributional thesaurus and sentence co-occurrences. Sarkar and Chakraborty (2015) also performed SA of Bengali tweets using SVM and MNB classifier. Se et al. (2015) reported the work on SA for Bengali tweets using machine learning classifiers and classified them into positive, negative and neutral class. The authors analyzed that machine learning classifiers perform better within constrained environment, i.e., without the availability of NLP tools such as POS tagger, NER. Ghosal et al. (2015) experimented on 6000 sentences of the Bengali horoscope corpus to perform SA. The authors used machine earning techniques such as NB, SVM, k-NN, DT and RF using features such as unigrams, bigrams and trigrams. The authors reported that SVM with 98.7% accuracy outperforms than other techniques without removing stop words and applying Information Gain (IG) as a feature selection method. Hassan et al. (2016) experimented using the deep recurrent model, namely LSTM

using two types of loss functions such as binary cross-entropy and categorical cross-entropy to perform SA for Bangla and Romanized Bangla text. The authors analyzed that categorical cross-entropy model performs better with 78% accuracy. Phani et al. (2016) used the annotated SAIL-2015 dataset (Patra et al. 2015) to perform SA for Bengali language. The authors experimented using six classifiers such as NB, LR, DT, RF, SVM using four categories of features, namely word n-grams, character n-grams, surface and SWN features. Seshadri et al. (2016) performed SA of Bengali tweets on a dataset of SAIL-2015 using RNN and classified tweets into positive, negative and neutral class.

(iii) **Tamil** Kumar et al. (2015c) performed SA of Tamil tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed SA using a regularized least square method. Se et al. (2015) reported the work on SA for Tamil tweets using machine learning classifiers and classified them into positive, negative and neutral class. The authors analyzed that machine learning classifiers perform better within constrained environment, i.e., without the availability of NLP tools such as POS tagger, NER. Nivedhitha et al. (2016) proposed an unsupervised dictionary based technique to perform SA of Tamil tweets. The authors used GENISM-Word2Vec topic modeling toolkit to convert the string data into vector form and performed sentiment classification using HSWN. Sharmista and Ramaswami (2016) experimented on 100 Tamil product reviews to perform SA using decision tree classification techniques such as J48, LMT, BagCart, Recursive, RF and C50. The authors attained an accuracy of 0.9469 and 0.9457 for LMT and random forest respectively. Se et al. (2016) performed SA of Tamil movie reviews using ML techniques such as NB, J48, SVM and ME. The authors also performed SA on the same dataset by considering SentiWordNet words as features and applied the four ML algorithm and concluded that SVM achieves best accuracy of 75.9% in comparison to other for SentiWordNet features. Seshadri et al. (2016) performed SA of Tamil tweets on a dataset of SAIL-2015 using RNN and classified tweets into positive, negative and neutral class. Phani et al. (2016) used the annotated SAIL-2015 dataset (Patra et al. 2015) to perform SA for Tamil language and experimented using six ML classifiers using different features, namely word n-grams, character n-grams, surface and SentiWordNet features.

(iv) **Malayalam** Anagha et al. (2014) proposed dictionary based approach to perform SA and also developed lexical resource file of 2000 sentiment words for the Malayalam text. The authors classified the Malayalam reviews into positive and negative classes by attaining an accuracy of 93.6%. Nair et al. (2014) proposed a rule-based approach to perform SA of Malayalam movie reviews at sentence level. The authors used Sandhi splitter for the tokenization of sentences and classified the sentences into three classes positive, negative and neutral. The authors also handled negations and smileys by building a dictionary pre-tagged with positive and negative sentiment. An accuracy of 85% was achieved. In 2015, the authors proposed hybrid approach using combination of rule-based and ML techniques. The authors also computed the ratings of reviews along with sentiment class. It was analyzed that SVM outperforms than Conditional Random Field (CRF) and achieved an accuracy of 91% (Nair et al. 2015). Jayan et al. (2015) proposed a hybrid approach by combining rule-based and ML to perform SA of Malayalam film reviews at aspect, sentence and document level. The authors used CRF for tagging of dataset and then applied rules to classify the documents into three classes such as positive, negative and neutral. Anagha et al. (2015) proposed a fuzzy based approach to perform SA of Malayalam movie reviews. The authors used TnT (Trigrams'n'Tags)

tagger to tag the input corpus and then fuzzy triangular membership function to extract the sentiment from text. The precision rate of 91.6% was reported, while comparing the system's output with manually tagged output. Thulasi and Usha (2016) performed aspect-based SA on Malayalam movie and product reviews and achieved an accuracy of 84.7%.

(v) **Urdu** Syed et al. (2010) proposed lexicon based approach to perform SA of Urdu text by manually creating sentiment lexicon. The authors extracted SentiUnits from text using shallow parsing. The authors further extended this work to handle the implicit negation problem and tested their system on the data set of movie reviews (Syed et al. 2011). Earlier, the authors worked on sentences representing single target. In this work, the authors extended their model to handle the presence of multiple targets as in the comparative sentences and used dependency parsing algorithm to associate the SentiUnits to their targets. The authors tested their modified approach on a dataset of movie reviews and electronic appliances and achieved an accuracy of 82.5% (Syed et al. 2014). Rehman and Bajwa (2016) used lexicon based approach to perform SA of Urdu news articles using Urdu SWN and achieved an accuracy of 66% by classifying the documents into positive and negative class. Mukhtar and Khan (2017) used ML approaches to perform SA of Urdu blogs. Mukhtar et al. (2017) validated their SA results using three standard evaluation measures, i.e., McNemar'stest, kappa statistic, and root mean squared error.

(vi) **Kannada** Deepamala and Kumar (2015) performed SA of Kannada documents. They manually created a polarity lexicon for Kannada language consisting of 5043 words and compared the accuracies of lexicon based approach with NB and ME. They observed that ME with 93% accuracy outperforms than lexicon based approach and NB. Kumar et al. (2015b) performed SA on Kannada web documents by exploring the usefulness of semantic and machine learning approaches. They identified that in the case of semantic approach, baseline method outperforms than other semantic approaches like negation, sentence based and Turney's methods. In case of ML approaches, NB performs better than other supervised learning methods such as DT, RF, Sequential minimal optimization (SMO), Abstract Data Type (ADT) Tree and Breadth First. The authors concluded that the precision of ML approaches is 7.22% better than semantic approaches. Hegde and Padma (2015) performed SA of Kannada mobile product reviews extracted from newspaper 'Prajavani' using lexicon based approach for aspect extraction and NB classifier to identify the polarity of reviews. The authors reported an accuracy of 65% but the system lacks in handling the multi class, comparative and conditional sentences. Rohini et al. (2016) performed feature based SA of Kannada movie reviews using decision tree. The authors extracted nouns as features and adjectives as sentiment words using Kannada POS Tagger. Hegde and Padma (2017) used Random Forest Ensemble after extending previous corpus of mobile product reviews and improved accuracy from 65% to 72% in this work.

8.2 Languages with minor research work

This sub section discusses the research work on Indian languages such as Punjabi, Oriya, Telugu, Nepali, Marathi, Konkani and Manipuri (belonging to Indian language families) which have contributed a little in the field of SA.

(i) **Punjabi** Kaur and Gupta (2014a) performed SA of Punjabi text using a hybrid approach using n-gram model and NB on a dataset collected from Newspapers and blogs. They compared their approach with existing approaches such as Hindi Subjective Lexicon,

HSWN, bilingual Dictionary, and Translated Dictionary. Kaur and Gupta (2014b) proposed an algorithm for SA of Punjabi text. They used a bilingual dictionary based approach to develop a subjective lexicon for Punjabi language using HSWN. The authors validated their approach using a subjective lexicon and analyzed that their approach is better over existing approaches. Arora and Kaur (2015) developed an offline application to perform SA of Punjabi political reviews using scoring approach.

(ii) **Oriya** Jena and Chandra (2014) performed opinion mining of Oriya text using SVM. Sahu et al. (2016a) performed SA on a dataset of 1000 sentences of movie reviews in Odia language using NB classifier and achieved an accuracy of 92%. Then the authors applied three supervised classification techniques such as NB, LR and SVM on a dataset of 6000 sentences and compared the performance of these techniques using evaluation measures precision, recall and accuracy. It was analyzed that LR with accuracy 88% outperformed than NB and SVM (Sahu et al. 2016b).

(iii) **Nepali** Gupta and Bal (2015) performed the first work on sentiment detection of Nepali text on a dataset of 25,435 sentences collected from online Nepali National Dailies, namely 'ekantipur' and 'nagariknews'. The authors developed their own Nepali SWN, namely 'Bhavanakos' and compared NB with resource based SWN approach. It was concluded that ML approach is better than resource based approach. Thapa and Bal (2016) reported work on SA for Nepali language on a dataset of 384 book and movie reviews at document level. The authors used Bag-of-words and Term Frequency (TF)-Inverse Document Frequency (IDF) features extraction models with and without stop words removal. The authors classified the reviews by applying classifiers such as SVM, Multinomial NB and LR. The authors compared the performance of classifiers with evaluation metrics such as F-measure and accuracy; and concluded that MNB outperforms than SVM and LR with any of feature extraction method.

(iv) **Marathi** Balamurali et al. (2012) used WordNet sense-based features to perform SA of travel reviews. Chaudhari et al. (2017) performed SA of Marathi documents using Gate Processor (Natural language processor) and Marathi WordNet to compute the sentiment polarity.

(v) **Telugu** Mukku et al. (2016) performed SA of Telugu sentences collected from Indian Languages Corpora Initiative (ILCI). The authors used Doc2Vec tool for converting sentences into sentence vectors and performed SA using ML techniques such as NB, LR, SVM, DT, MLP Neural Network and RF. The authors also experimented by an ensemble of all the six ML classifiers. Naidu et al. (2017) proposed a two-phase SA for Telugu news sentences using Telugu SentiWordNet. First, the authors performed subjectivity classification then further classified them into positive and negative sentences.

(vi) **Konkani** Miranda and Mascarenhas (2016) developed an opinion mining system for Konkani language, namely KOP (Konkani OPinion mining system). The authors used Konkani SWN to perform SA and also handled negations, conjunctions as well as sarcasm.

(vii) **Manipuri** Nongmeikapam et al. (2014) performed SA on Manipuri text collected from daily newspapers. They processed the text for POS tagging using CRF then identified the verbs using a modified verb lexicon. After that, they counted the polarity for each class, such as positive, negative and neutral separately. Then highest polarity of the three decided the sentiment polarity of the document.

Table 7 summarizes the different approaches, corpora, corpus sizes, lexical resources/tools/programming languages and evaluation measures used to develop SA systems for all the Indian languages considered in this study.

Table 7 Summary about approaches and lexical resources for different Indian languages

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Konkani (Miranda and Mascarenhas 2016)	Lexicon Based	General	Not specified	SWN	Not specified
Oriya (Jena and Chandra 2014)	SVM	General	Not specified	SVM library	Not specified
Odia (Sahu et al. 2016a)	NB	Movie	1000 sentences	Natural Language Toolkit (NLTK)	A: 0.92, P: 0.93, R: 0.97
Odia (Sahu et al. 2016b)	NB, SVM, Logistic Regression	Movie Reviews	6000 reviews	Python language	LR- P: 0.75, R:0.797, A: 0.88
Nepali (Gupta and Bal 2015)	Lexicon based	Nepali National Dailies	25435 sentences	Nepali SWN	P: 47.2, R:54.8
	NB				P: 23.6, R:70.2
Nepali (Thapa and Bal 2016)	SVM, MNB and LR	Book and Movie Reviews	384	Natural Language Toolkit and Python language packages	MNB- F: 0.67, A: 0.65
Manipuri (Nongmeikapam et al. 2014)	CRF	Newspapers	550 letters	POS Tagger	R: 72.10%, P: 78.14%, F: 75.00%.
Marathi (Balamurali et al. 2012)	SVM	Travel reviews	150 reviews	WordNet, LibSVM package	A: 84%
Marathi (Chaudhari et al. 2017)	NLP based	General	Not specified	Marathi WordNet, GATE	Not specified
Telugu (Mukku et al. 2016)	NB, SVM, DT, RF, MLP Neural Network, LR	Newspapers	1644 annotated sentences + 721,785 raw sentences	Python	Ensemble - A: 73.85% (binary class), A: 60.13% (ternary class)
Telugu (Naidu et al. 2017)	Lexicon based	Newspapers	1400 sentences	SWN	A: 81%, P:0.71, R: 0.77, F:0.74
Punjabi (Kaur and Gupta 2014a)	Hybrid (N-gram, NB)	Newspapers and Blogs	44,200 sentences	WEKA, Java	P:0.78, R:0.6, F: 0.67

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Punjabi (Kaur and Gupta 2014b)	Lexicon based	General	Not specified	SWN	A: 78.02
Punjabi (Arora and Kaur 2015)	Scoring approach	General	Not specified	SWN	Not specified
Kannada (Hegde and Padma 2015)	NB	Mobile product reviews	Not specified	Python language	A: 65%, P:62.5%, R: 75%, F: 68.2%
Kannada (Deepamala and Kumar 2015)	Lexicon based	General	344 documents	Kannada stemmer	ME- A: 0.93, P: 0.9, R:0.89, F:0.89
	NB, ME				
Kannada (Kumar et al. 2015b)	Lexicon based	Products reviews	197 reviews	Kannada POS Tagger	NB-P:0.81
	J48, Random Tree, ADT Tree, Breadth First, NB and SVM			WEKA	
Kannada (Hegde and Padma 2017)	RF	Mobile Products reviews	Not specified	R Studio	A:72%
Kannada (Rohini et al. 2016)	DT	Movie reviews	100 reviews	Kannada POS Tagger	P: 0.78, R:0.79
Malayalam (Anagha et al. 2014)	Lexicon based	Multi-domain reviews	Not specified	TnT Tagger, Malayalam lexical resource file	A: 93.6%
Malayalam (Nair et al. 2014)	Lexicon based	Movie Reviews	Not specified	Sandhi Splitter, Python	A: 85%
Malayalam (Nair et al. 2015)	Hybrid ([SVM, CRF] + rule based)	Movie Reviews	30,000 tokens	SVM and CRF libraries	SVM: P: 0.806, R: 0.951 and F: 0.863
Malayalam (Jayan et al. 2015)	Hybrid (CRF + rule based)	Movie reviews	30,000 tokens	CRF library	A:82%
Malayalam (Anagha et al. 2015)	Fuzzy logic	Movie Reviews	2500 words	TnT Tagger	P: 91.6%

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Malayalam (Thulasi and Usha 2016)	Lexicon based	Movie and Product reviews	50 sentences	Sandhi splitter, TnT Tagger, Malayalam SentiWordNet	A: 84.7%
Urdu (Syed et al. 2010)	Lexicon based	Movie and Product reviews	753 reviews	Shallow parser	Movie- A: 72%, Product- A: 78%
Urdu (Syed et al. 2011)	Lexicon based	Movie reviews	450 reviews	Shallow parser	Set1- P:0.864, R: 0.837, F: 0.850; Set2- P:0.590, R: 0.779, F: 0.677;Set3- P:0.510, R:0.615, F:0.558
Urdu (Syed et al. 2014)	Lexicon based	Movie and electronic appliances reviews	700 movie reviews, 650 electronic appliance reviews	Shallow parser	A: 82.5%
Urdu (Rehman and Bajwa 2016)	Lexicon based	News	124 documents	SWN	A: 0.66. R: 0.79, P: 0.69, F: 0.73
Urdu (Mukhtar and Khan 2017)	SVM, DT, k-NN	Blogs (14 domains)	6025 sentences	WEKA	k-NN- A: 67.0185%, P: 0.674, R: 0.6703, F: 0.6703
Tamil (Nivedhitha et al. 2016)	Lexicon based	Tweets	691 tweets	SWN, Genism Python toolkit	A: 0.7062, P: 0.7065, R:0.6987, F:0.6924
Tamil (Sharmista and Ramaswami 2016)	Decision tree classification techniques (J48, LMT, BagCart, Recursive, RF and C50)	Product reviews	100 reviews	R	LMT- A: 0.9469
Tamil (Se et al. 2016)	SVM, DT, NB, ME	Movie Reviews	534 reviews	SWN	SVM- A:75.9%
Tamil (Seshadri et al. 2016)	RNN	Tweets	SAIL-2015	MIKE (Mining Intelligence and Knowledge Exploration)	A: 88.23, F:0.802

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Tamil (Phani et al. 2016)	MNB, DT, SVM, RF, LR	Tweets	SAIL-2015	SWN	NB- A: 62.16% (2-class); RF-A: 45.24% (3-class)
Tamil (Kumar et al. 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 32.32%
Tamil (Se et al. 2015)	NB	Tweets	SAIL-2015	SciPy	Constrained- A: 39.28%
Bengali (Das and Bandyopadhyay 2010a)	Hybrid (SVM + rule based)	News	447 sentences	Dependency parser, SWN	P: 70.04%, R:63.02%
Bengali (Hasan et al. 2014)	Lexicon based	General	approx. 150 sentences	WordNet, SWN	Percentage of positity, negativity and neutrality
Bengali (Ghosal et al. 2015)	NB, SVM, k-NN, DT, Random Forest	Horoscopes	6000 sentences	WEKA	SVM- A:98.7%
Bengali (Hassan et al. 2016)	LSTM	General	6698 entries	Python's Kera library	A: 78%
Bengali (Phani et al. 2016)	MNB, DT, SVM, RF, LR	Tweets	SAIL-2015	SWN	NB- A: 67.83% (2-class); LR-A: 51.25% (3-class)
Bengali (Seshadri et al. 2016)	RNN	Tweets	SAIL-2015	MIKE	A: 65.16, F:0.644
Bengali (Kumar et al. 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 31.4%
Bengali (Kumar et al. 2015a)	SVM	Tweets	SAIL-2015	SWN	A: 43.2% (constrained); A: 42% (unconstrained)

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Bengali (Sarkar and Chakraborty 2015)	MNB	Tweets	SAIL-2015	WEKA, SWN	A: 41.2% (constrained); A: 40.4% (unconstrained)
Bengali (Se et al. 2015)	NB	Tweets	SAIL-2015	SciPy	A: 33.6% (constarined)
Hindi (Joshi et al. 2010)	Lexicon based	Movie Reviews	250 Hindi reviews, 1000 English Reviews	SWN	A: 60.31%
	Machine learning (SVM)			Rapid Miner 5.0	A: 78.14%
	Translation Based (SVM)			Google Translator	A: 65.96%
Hindi Pandey and Govilkar (2015)	Lexicon based	Movie Reviews	Not specified by the author	SWN	Not specified
Hindi (Sharma et al. 2015)	Lexicon based	Tweets	100	SWN	A: 77.75%, P: 0.85. R: 0.88
Hindi (Sharma and Bhattacharyya 2014)	Lexicon based	Movie and product reviews	900 reviews	SWN	A: 87%
Hindi (Akhtar et al. 2016b)	SVM	Movie, product, travel and mobile apps reviews	5417 reviews	Shallow Parser	A: 54.05%
Hindi (Akhtar et al. 2016a)	NB, DT and SMO	Movie, product, travel and mobile apps reviews	5254 reviews	WEKA	DT- A: 54.48% (Products), A: 47.95% (Mobile apps), A: 65.20% (Travels), A: 91.62% (Movies)

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Hindi (Seshadri et al. 2016)	RNN	Tweets	SAIL-2015	MIKE	A: 72.01, F:0.714
Hindi (Jha et al. 2015)	NB	Movie reviews	200 reviews	NLTK	A: 87.1%
	Lexicon based			TnT POS Tagger	
Hindi (Sharma et al. 2014)	Lexicon based	Movie Reviews	Not specified	POS tagger	A: 0.65, P: 0.66, R: 0.78
Hindi (Bansal et al. 2013)	Deep belief Network	Movie Reviews	300 reviews	Theano Library	A: 64%
Hindi (Phani et al. 2016)	MNB, DT, SVM, RF, Logistic Regression	Tweets	SAIL-2015	SWN	LR- A: 81.57% (2-class); LR-A: 56.96% (3-class)
Hindi (Mittal et al. 2013)	Lexicon based	Movie Reviews	662 reviews	SWN	A: 80.21%
Hindi (Arora 2013)	Lexicon based, N-gram Modeling	Products and Movie Reviews	973 reviews	WEKA	A: 61.6% (N-gram + lexical features)
Hindi (Sharma and Moh 2016)	Lexicon Based	Election tweets	42,235 tweets	SWN	A: 34%
	NB, SVM				NB- A: 62.1%, P:0.71, R:0.61
Hindi (Bakliwal et al. 2012)	Lexicon based	Products reviews	700 reviews	Hindi Shallow parser, Hindi subjective lexicon	A: 79.03%
Hindi (Prasad et al. 2015)	DT	Tweets	SAIL-2015	WEKA	Constrained- P: 0.822, R: 0.815, F: 0.804; Unconstrained- P: .735, R: .707, F: 0.680

Table 7 continued

Language (Author)	Approach	Corpus type	Corpus size	Resources used	Evaluation measures
Hindi (Venugopalan and Gupta 2015)	SVM, DT	Tweets	SAIL-2015	WEKA	SVM- A:42.83%
Hindi (Kumar et al. 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 47.96%
Hindi (Kumar et al. 2015a)	SVM	Tweets	SAIL-2015	SWN	Constrained- A: 49.68%, Unconstrained- A: 46.25%
Hindi (Sarkar and Chakraborty 2015)	MNB	Tweets	SAIL-2015	WEKA, SWN	Constrained- A: 50.75 %, Unconstrained- A: 48.82 %
Hindi (Se et al. 2015)	NB	Tweets	SAIL-2015	SciPy	Constrained-A: 55.67%
Hindi (Balamurali et al. 2012)	SVM	Travel reviews	200 reviews	WordNet, LibSVM package	A: 72%
Hindi (Akhtar et al. 2016c)	Hybrid	Tweets, Products, Movies, restaurant reviews	2152 reviews	DL4J, LibSVM package	Tweets-A: 58.62, Laptop reviews-A: 68.04. Restaurant reviews-A: 77.16

P: precision, R: recall, A: accuracy, F: F-measure

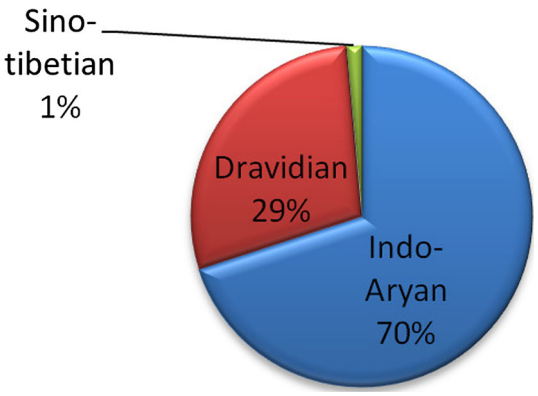

Fig. 13 Status of SA research work in Indian language families

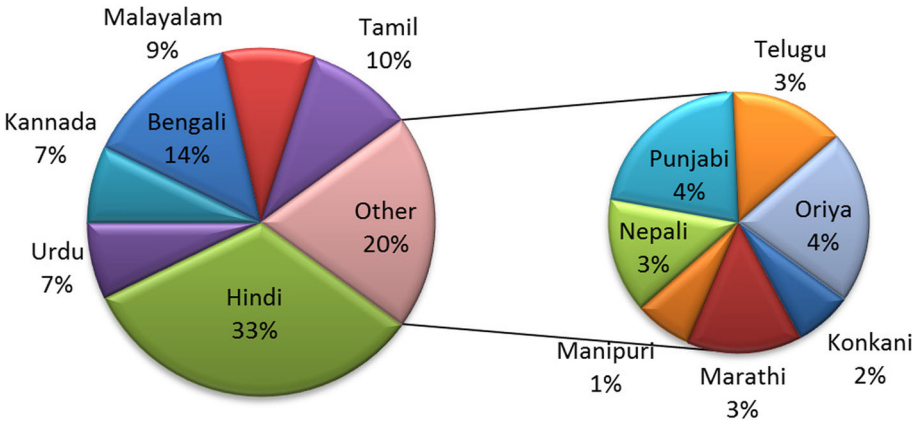

Fig. 14 Status of SA research work in different Indian languages

9 Findings of systematic survey

This section concludes the results identified while conducting this systematic survey and efforts have been made to answer all the research questions given in Table 1. The answer of research question **RQ3** is reported through Figs. 13 and 14 which show the percentage of research studies covering different Indian languages under Indian language families over a period from 2010 to 2017. From Fig. 13, it can be observed that 70% of the research work on SA has been performed on Indo-Aryan languages (out of which major part, i.e., 33% is covered by Hindi language), followed by 29% on Dravidian and 1% on Sino-Tibetian language families. And Austroasiatic language family is still unexplored for SA research work. Figure 14 depicts that the majority of research has been done for Hindi language (33%), followed by Bengali (14%), Tamil (10%), Malayalam (9%), Urdu (7%), Kannada (7%), Punjabi (4%), Oriya (4%), Nepali (3%), Telugu (3%), Marathi (3%), Konkani (2%) and Manipuri (1%). The systematic map in Fig. 14 helps in recognizing the mostly explored Indian languages in the field of sentiment analysis from 2010 to 2017.

From this analysis, it is concluded that one-third of the research work has been done on Hindi language belonging to the Aryan languages family. However, with the introduction of Unicode (UTF-8) standards, web pages in Hindi language have been increasing rapidly. Hindi is spoken by a total of 422 million people; it's about

41% of total population of India. The government of India is also promoting the Hindi language by providing online contents of official websites in Hindi. Therefore, researchers are attracting towards performing SA in Hindi language so that the large volume of opinions shared by people on web can be effectively leveraged.

The answer for the research question **RQ4** is attained through Table 8 which summarizes the SA techniques used by researchers for all Indian languages. It reports the SA research work for Indian languages along with publications count over which different SA techniques have been experimented.

From Table 8, it can be stated that lexicon based approach has been experimented over almost all Indian languages which concludes that researchers first experimented with SA in their own language by constructing polarity lexicons, while majority of the researchers have used ML approaches followed by lexicon based, deep learning and hybrid as shown in Fig. 15a. There is too much potential in machine learning, overtaking some of the manual labor of some lexicon based tasks that are labor intensive. For example, lexicon sentiment creation is labor intensive. Therefore, majority of the research studies(i.e., approx. 60%) opted for ML to perform SA. However, in recent times, researchers are also attracting towards experimenting with deep learning techniques due to improvement in accuracy irrespective of time constraint that is needed to train the data. Out of ML approaches researchers mostly used SVM, NB and DT as shown in Fig. 15b covering approximately 70% of the research studies for the development of SA systems. Figure 15c shows that in case of lexicon-based techniques, 60% of the researchers have created their own SWN using bi-lingual dictionary based approach and 25% of the researchers started with some seed list of polarity words and used Wordnet based approach to extend the sentiment lexicon while remaining 15% of the researchers manually developed SWN for their own language.

The answer to research question **RQ5** has been addressed through Fig. 16 which depicts the domains considered by the researchers to perform SA for different Indian languages. It has been observed that mainly the research work has been performed on movie reviews and tweets. The reason behind the majority of research work on movie reviews is due to availability of annotated dataset. In recent years, researchers are performing SA on social media sites such as Twiiter and Facebook to analyze the sentiments of people real-time. One example of social media site acting as valuable resource is that during the 2009 Jakarta and Mumbai terrorist attacks, Twitter played a vital role to harvest civilian statement and response. Due to these scenarios, researchers are attracting towards SA of real time data such as tweets, Facebook posts etc.

Table 9 summarizes the research studies according to the different sentiment levels such as aspect, sentence and document; classes namely positive, negative and neutral; and whether the SA system handles negations or not. It is analyzed that approximate 30% of the research studies have handled negations which also plays a major role in improving sentiment classification. Table 9 also assists in concluding that mostly the researchers (i.e., 72%) have worked at sentence level. The remaining 28% of the researchers cover the SA work on document level as well as aspect-level as shown in Fig. 17a which states that the research work on aspect-level is still in growing phase. As the SA at aspect-level helps in performing fine-grained analysis therefore the researchers are attracting towards it. As shown in Fig. 17b, majority of researchers (i.e., 64%) have considered have only 2 classes, i.e., positive and negative for SA of Indian languages despite of considering the third "neutral" class. The reason behind

Table 8 Indian languages for each SA technique

Techniques	Languages (Count)	References
Lexicon based	Hindi (9), Urdu (5), Malayalam (3), Kannada (2), Konkani (1), Nepali (1), Bengali (1), Punjabi (1), Tamil (1), Telugu (1)	Bakliwal et al. (2012); Miranda and Mascarenhas (2016); Gupta and Bal (2015); Anagha et al. (2014); Nair et al. (2014); Thulasi and Usha (2016); Deepamala and Kumar (2015); Kaur and Gupta (2014b); Hasan et al. (2014); Syed et al. (2010, 2011); Joshi et al. (2010); Nivedhitha et al. (2016); Pandey and Govilkar (2015); Sharma et al. (2015); Sharma and Bhattacharyya (2014); Naidu et al. (2017); Rehman and Bajwa (2016); Jha et al. (2015); Sharma et al. (2014); Mittal et al. (2013); Sharma and Moh (2016)
SVM	Hindi (7), Bengali (3), Oriya (2), Tamil (2), Nepali (1), Telugu (1), Kannada (1), Marathi (1), Urdu (1)	Jena and Chandra (2014); Sahu et al. (2016b); Thapa and Bal (2016); Mukku et al. (2016); Kumar et al. (2015b); Phani et al. (2016); Ghosal et al. (2015); Joshi et al. (2010); Venugopalan and Gupta (2015); Kumar et al. (2015a); Akhtar et al. (2016b, a); Balamurali et al. (2012); Mukhtar and Khan (2017); Se et al. (2016); Sharma and Moh (2016)
NB	Hindi (3), Kannada (3), Oriya (2), Bengali (2), Tamil (2), Nepali (1), Telugu (1), Urdu (1)	Sahu et al. (2016a, b); Gupta and Bal (2015); Mukku et al. (2016); Hegde and Padma (2015); Deepamala and Kumar (2015); Kumar et al. (2015b); Ghosal et al. (2015); Se et al. (2015); Akhtar et al. (2016a); Mukhtar and Khan (2017); Se et al. (2016); Jha et al. (2015); Sharma and Moh (2016)
DT	Hindi (4), Kannada (2), Bengali (2), Tamil (2), Tamil (1), Urdu (1), Telugu (1)	Mukku et al. (2016); Kumar et al. (2015b); Phani et al. (2016); Ghosal et al. (2015); Prasad et al. (2015); Venugopalan and Gupta (2015); Akhtar et al. (2016a); Rohini et al. (2016); Mukhtar and Khan (2017); Sharmista and Ramaswami (2016); Se et al. (2016)

Table 8 continued

Techniques	Languages (Count)	References
Logistic Regression	Oriya (1), Nepali (1), Telugu (1), Bengali (1), Hindi (1), Tamil (1)	Sahu et al. (2016b); Thapa and Bal (2016); Mukku et al. (2016); Phani et al. (2016)
Random Forest	Bengali (2), Tamil (2), Telugu (1), Hindi (1), Kannada (1)	Mukku et al. (2016); Phani et al. (2016); Ghosal et al. (2015); Hegde and Padma (2017); Sharmista and Ramaswami (2016)
MNB	Hindi (2), Bengali (2), Tamil (1), Nepali (1)	Thapa and Bal (2016); Phani et al. (2016); Sarkar and Chakraborty (2015)
ME	Kannada (1), Tamil (1)	Deepamala and Kumar (2015); Se et al. (2016)
k-NN	Bengali (1)	Ghosal et al. (2015)
MLP Neural Network	Telugu (1)	Mukku et al. (2016)
Hybrid	Malayalam (2), Hindi (1), Punjabi (1), Bengali (1)	Akhtar et al. (2016c); Nair et al. (2015); Jayan et al. (2015); Kaur and Gupta (2014a); Das and Bandyopadhyay (2010a)
Deep learning	Bengali (2), Hindi (2), Tamil (1)	Hassan et al. (2016); Seshadri et al. (2016); Bansal et al. (2013)
Fuzzy logic	Malayalam (1)	Anagha et al. (2015)
CRF	Manipuri (2)	Nongmeikapam et al. (2014)
Regularized Least Square	Tamil (1), Bengali (1)	Kumar et al. (2015c)
NLP based	Marathi (1)	Chaudhari et al. (2017)
Scoring based	Punjabi (1)	Arora and Kaur (2015)

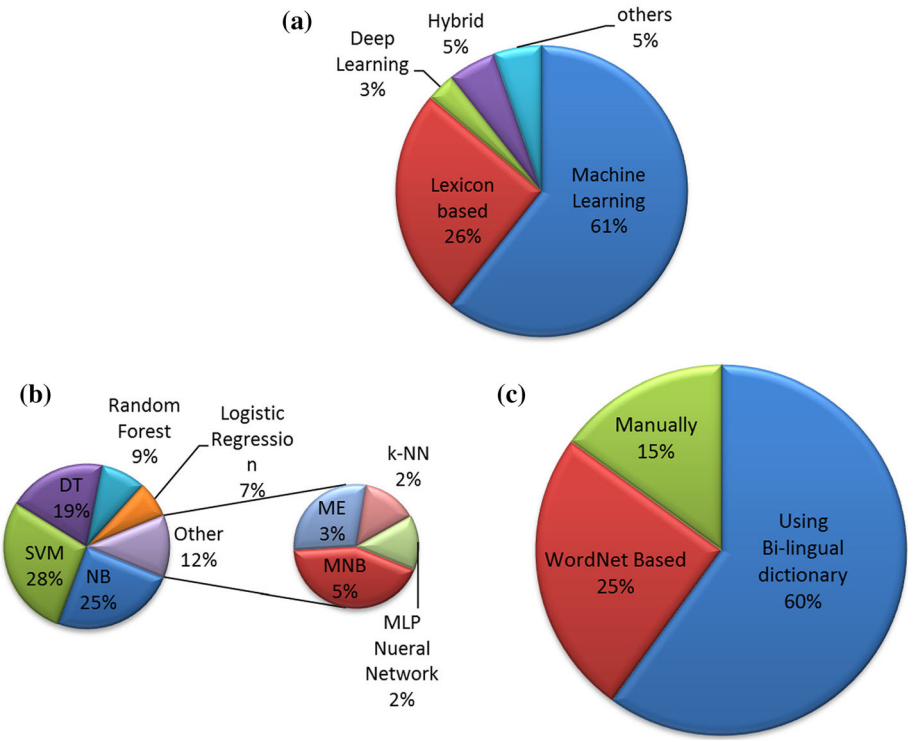

Fig. 15 Percentage of research work using **a** SA techniques, **b** ML techniques, and **c** Lexicon-based techniques

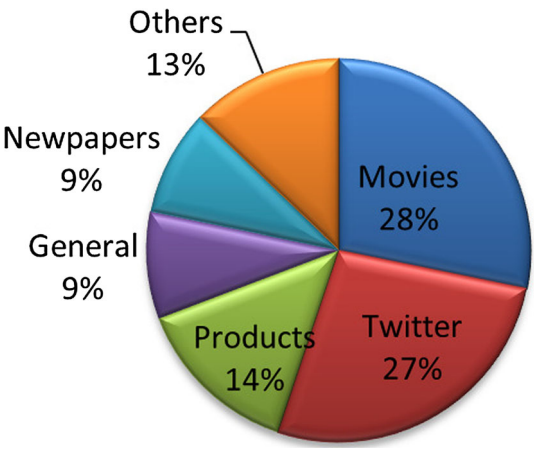

Fig. 16 Percentage of SA work for different domains

consideration of 2 classes for SA process is that better accuracy can be achieved for 2 classes in comparison to three classes. Figure 17a, b help to find the answer for research question **RQ6**.

Table 9 Classification of Indian languages according to different parameters

Language (Author)	Aspect	Sentence	Document	Positive	Negative	Neutral	Handling of negations
Odia (Sahu et al. 2016a; Akhtar et al. 2016c), Nepali (Gupta and Bal 2015; Mukku et al. 2016), Marathi (Balamurali et al. 2012), Telugu (Naidu et al. 2017), Urdu (Syed et al. 2014), Tamil (Se et al. 2016), Bengali (Ghosal et al. 2015; Hassan et al. 2016), Hindi (Joshi et al. 2010; Balamurali et al. 2012; Bansal et al. 2013)	–	+	–	+	+	–	–
Telugu (Mukku et al. 2016), Malayalam (Anagha et al. 2015), Urdu (Rehman and Bajwa 2016; Mukhtar and Khan 2017), Tamil (Kumar et al. 2015c; Nivedhitha et al. 2016; Sharmista and Ramaswami 2016; Se et al. 2016), Bengali (Phani et al. 2016; Kumar et al. 2015c, a; Sarkar and Chakraborty 2015; Se et al. 2015; Seshadri et al. 2016), Hindi (Phani et al. 2016; Prasad et al. 2015; Venugopalan and Gupta 2015; Kumar et al. 2015c, a; Sarkar and Chakraborty 2015; Se et al. 2015; Seshadri et al. 2016)	–	+	–	+	+	+	–
Punjabi (Kaur and Gupta 2014b), Malayalam (Nair et al. 2014), Hindi (Bakliwal et al. 2012; Pandey and Govilkar 2015; Sharma et al. 2015, 2014; Arora 2013; Sharma and Moh 2016)	–	+	–	+	+	+	+
Marathi (Chaudhari et al. 2017), Malayalam (Anagha et al. 2014), Bengali (Hasan et al. 2014)	–	+	+	+	+	+	–
Malayalam (Nair et al. 2015), Hindi (Sharma and Bhattacharyya 2014)	–	+	–	+	+	–	+
Oriya (Jena and Chandra 2014), Manipuri (Nongmeikapam et al. 2014)	–	–	+	+	+	+	–
Kannada (Deepamala and Kumar 2015), Urdu (Syed et al. 2010)	–	+	+	+	+	–	+
Punjabi (Kaur and Gupta 2014a), Kannada (Kumar et al. 2015b)	–	+	+	+	+	+	+
Malayalam (Thulasi and Usha 2016), Hindi (Akhtar et al. 2016a, b)	+	–	–	+	+	+	–
Kannada (Hegde and Padma 2015, 2017; Rohini et al. 2016)	+	+	–	+	+	–	–

Table 9 continued

Language (Author)	Aspect	Sentence	Document	Positive	Negative	Neutral	Handling of negations
Malayalam (Jayan et al. 2015)	+	+	+	+	+	+	–
Konkani (Miranda and Mascarenhas 2016)	+	+	+	+	+	+	+
Punjabi (Arora and Kaur 2015)	–	+	+	+	+	–	–
Bengali (Das and Bandyopadhyay 2010a)	+	–	–	+	+	–	+
Hindi (Mittal et al. 2013)	–	–	+	+	+	–	+
Hindi (Jha et al. 2015)	–	–	+	+	+	+	+
Urdu (Syed et al. 2011)	+	+	–	+	+	–	+
Hindi (Akhtar et al. 2016c)	+	+	–	+	+	+	–

'+' indicates that corresponding language supports corresponding factor that are used for SA

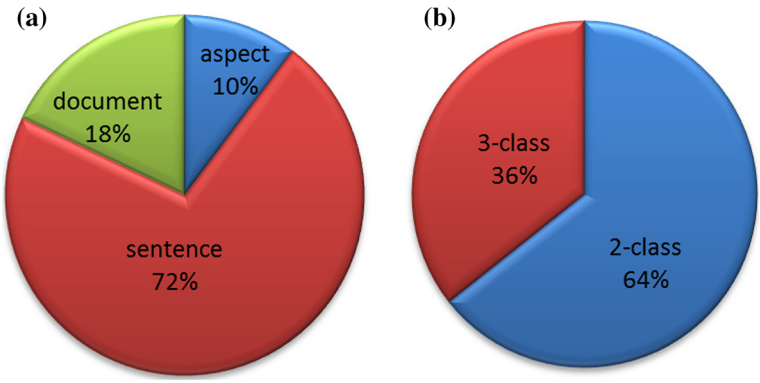

Fig. 17 Percentage of work **a** at different sentiment levels, **b** for different sentiment classes

To address the research question **RQ7**, the online available SA tools such as Alchemy API,[1] Semantria,[2] Trackur,[3] Sentigem,[4] etc. have been explored, it has been identified that till now, no system has been built which online performs SA of Indian languages. The answer to research question **RQ8** is addressed in Sect. 10.

The major findings of the research questions listed in Table 1 from this systematic survey can be summarized as follows.

- The research work in the field of SA for Indian languages started in 2010 when Joshi et al. (2010) first set the benchmark by performing SA for Hindi language. This research study has highest impact in the field of SA over Indian languages till now as it has more than fifty citations.
- After 2010, it has been observed that SA research work has been performed in 13 Indian languages such as Hindi, Bengali, Tamil, Malayalam, Kannada, Urdu, Punjabi, Oriya, Nepali, Telugu, Konkani, Manipuri and Marathi out of 22 languages.
- This research study provides the brief description about the different SA techniques such as ML, lexicon based and deep learning.
- This systematic review helps in providing knowledge about the online availability of annotated datasets, linguistics resources and polarity lexicons for different Indian languages.
- The annotated datasets are available for Hindi, Bengali, Tamil and Marathi. Similarly, linguistic resources such as Morph analyzer, POS tagger, dependency parser are available for different Indian languages. Researchers can easily use these resources as description along with online availability is provided in this systematic review.
- From this survey, it has been observed that majority of research work in the field of SA has been performed for Indo-Aryan language, i.e., Hindi that covers approximate one-third of the research work performed for Indian languages.
- It has been analyzed that mostly the researchers have used ML techniques, however the researchers are also attracting towards deep learning techniques due to better accuracy achieved by these techniques.

[1] http://www.alchemyapi.com/.

[2] https://www.lexalytics.com/.

[3] http://www.trackur.com/.

[4] http://sentigem.com/.

– From this systematic survey, it can be observed that researchers have performed mostly SA work on sentence level for positive and negative sentiment classes. Also, mainly the authors have experimented on movie reviews dataset and tweets.

It has also been observed that an extensive amount of research work has been done for English language using different approaches and techniques. Therefore, these research results can also be transferred to Indian languages. To perform the SA using lexicon-based approach for Indian languages, one can develop the own sentiment lexicon by using resources English SWN and Indo WordNet. For this, one can map the synsets of English SWN along with polarity scores to the synsets of target language with the help of Indo WordNet. Also, the best performing ML technique used by researchers for the SA of English language can also be applied on Indian language dataset but it requires the annotated dataset for the target language. However, accuracy may be decreased due to the differences in English and Indian language structure as discussed in Sect. 4.3.

10 Conclusions and future work

The growth of research work in the field of SA for Indian language content motivated us to conduct this systematic survey. In India, there are 22 official languages and due to availability of data from multiple sources for each language, it is easy to gather data and analyze them. The research work on SA in context to Indian languages was first commenced by Joshi et al. (2010), the highest cited research study till now. Afterwards it, the research work in this field is continuously growing from last couple of years as Indian language content on web is also increasing. Till now, no research study is available which covers an in-depth analysis for Indian languages in the field of SA. Therefore, this paper is a significant contribution in the literature of SA for Indian languages which includes the systematic survey over 59 research studies published on SA for all Indian language families from 2010 to 2017 (till end of July) to include the relevant work only. However, with the pace of research and development in SA field, this paper can be further extended by including the recent and upcoming SA studies (Mukhtar et al. 2018a; Rani and Kumar 2018a, b) which have performed SA for different Indian languages. The 59 research studies considered in this systematic survey have been decided by developing a review protocol which includes the research questions, sources of information, inclusion and exclusion criteria. The different findings of this survey have been analyzed to get the answers of the targeted research questions framed in this paper.

The summary about the different SA approaches, type and size of corpora, lexical resources/tools and evaluation measures for each Indian language is given in this paper. From this summary, it has been analyzed that SA work has been reported on 13 Indian languages and majority of the work in this field has been published in conferences followed by journals. It has been observed from the comprehensive analysis that 70% of the research work has been done for Indo-Aryan language family in which major part is covered by Hindi and Bengali language (i.e., 47%). It has also been noticed that the researchers have mainly used ML (i.e., 61%) approach in comparison to other lexicon based, deep learning and hybrid approaches. Also, the researchers have performed mainly SA work at sentence level and considered two sentiment classes, i.e., positive and negative in majority of research studies using different domains like tweets, movies and products reviews etc. This paper also gives the details about online available annotated datasets, pre-processing linguistic resources available for different Indian languages which can help the researchers to perform SA in other Indian languages.

The online available SWN(s) for various Indian languages and the approaches to develop them are also discussed in this paper.

Also, there is a large amount of work that needs to be done to understand the behavior of Indian languages and to improve the accuracy of the SA systems. It has been observed that researchers don't make available their annotated datasets online. As there is lack of annotated datasets and creation of labeled datasets for different Indian languages is a time consuming task. In future, the annotated resources can be made available for utilizing these by other researchers so that they can focus only on improving the accuracy of the system by developing new SA approaches. It has also been analyzed that deep learning approaches for SA are in demand. Therefore, researchers can experiment with these approaches to achieve better results in future. The basic linguistic resources like shallow parsers, dependency parsers, POS taggers and morphological analyzers need to be improved to enhance the accuracy of the lexicon based approaches. Also, the existing SWN(s) available for different Indian languages are required to be improved by making these SWN(s) dynamic. Also, there is a need to work on the trending code-mixed data as a little work has been done in this area. By improving the SA approaches, the research work can be carried out in remaining Indian languages and also the accuracy of the existing SA systems can be improved. And, also there is a need to build online systems which can perform SA for Indian languages. The use of punctuation is a hurdle in the area of SA which is under research as well.

The identification and handling of intensifiers in SA process can also help in improvement of its performance (Mukhtar et al. 2018b). Emotions play an important role and detection of emotions is considered as SA task which aims to detect various types of emotions such as joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. The existing research work has not dealt with emotion detection till now and it can be performed in future which can help in improving SA process. Now a days, people express their sentiments with images and videos along with text on social media platforms such as Facebook, Instagram, etc. Sentiment analysis of images and videos will have to pace up with this change. Sometimes, people use emoticons while typing messages, statuses on Facebook, Whatsapp etc., therefore a SA system can be developed in future which can automatically detect the emoticon by anlyzing the sentiment from text. A few authors (Asghar et al. 2018b) have developed a framework to detect the slangs and emotions for English language for new text types like tweets, the same techniques need to be adapted for Indian languages to pre-process as well as to tackle the informal style of the target language to obtain acceptable levels of performance of the SA systems. This field needs to be combined with effective computing, psychology and neuroscience to converge on a unified approach to understand the sentiments better. Thus, this survey can help the researchers in building the effective SA for their own Indian language by using the different methods and techniques used by other researchers which can help in benefits of society.

Acknowledgements This Publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

(2012) Shallow parsers, Language Technologies Research Centre (LTRC), IIIT Hyderabad. http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php. Accessed 25 June 2017

(2014) Indo-Aryan languages. http://www.indianetzone.com/11/indo_aryan_languages.htm. Accessed 22 June 2017

(2015) Indian language families. http://www.indianetzone.com/39/indian_language_families.htm. Accessed 20 June 2017
(2017) Online education in India: 2021. https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/05/Online-Education-in-India-2021.pdf. Accessed 15 June 2017
Akhtar MS, Ekbal A, Bhattacharyya P (2016a) Aspect based sentiment analysis: category detection and sentiment classification for Hindi. In: 17th International conference on intelligent text processing and computational linguistics, pp 1–12
Akhtar MS, Ekbal A, Bhattacharyya P (2016b) Aspect based sentiment analysis in Hindi: resource creation and evaluation. In: Proceedings of the 10th international conference on language resources and evaluation, pp 1–7
Akhtar MS, Kumar A, Ekbal A, Bhattacharyya P (2016c) A hybrid deep learning architecture for sentiment analysis. In: Proceedings of the 26th international conference on computational linguistics, pp 482–493
Anagha M, Kumar RR, Sreetha K, Rajeev R, Raj PR (2014) Lexical resource based hybrid approach for cross domain sentiment analysis in Malayalam. Int J Eng Sci 15:18–21
Anagha M, Kumar RR, Sreetha K, Raj PR (2015) Fuzzy logic based hybrid approach for sentiment analysisl of malayalam movie reviews. In: International conference on signal processing. Informatics, communication and energy systems. IEEE, pp 1–4
Arora P (2013) Sentiment analysis for Hindi language. MS by Research in Computer Science
Arora P, Kaur B (2015) Sentiment analysis of political reviews in Punjabi language. Int J Comput Appl 126(14):1–4
Asghar MZ, Khan A, Zahra SR, Ahmad S, Kundi FM (2017) Aspect-based opinion mining framework using heuristic patterns. Clust Comput. https://doi.org/10.1007/s10586-017-1096-9
Asghar MZ, Khan A, Khan F, Kundi FM (2018a) Rift: a rule induction framework for twitter sentiment analysis. Arab J Sci Eng 43(2):857–877
Asghar MZ, Kundi FM, Ahmad S, Khan A, Khan F (2018b) T-saf: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Syst 35(1):1–19
Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Proc Lang Resour Eval 10:2200–2204
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, pp 1189–1196
Balamurali A, Joshi A, Bhattacharyya P (2012) Cross-lingual sentiment analysis for Indian languages using linked Wordnets. In: Proceedings of 24th international conference on computational linguistics: posters, pp 73–82
Bansal N, Ahmed UZ, Mukherjee A (2013) Sentiment analysis in Hindi. Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India, pp 1–10
Bhattacharyya P (2017) Indowordnet. In: The WordNet in Indian languages. Springer, pp 1–18
Chand S (2016) Indian languages: classification of Indian languages. http://www.yourarticlelibrary.com/language/indian-languages-classification-of-indian-languages/19813/. Accessed 22 June 2017
Chaudhari CV, Khaire AV, Murtadak RR, Sirsulla KS (2017) Sentiment analysis in Marathi using Marathi WordNet. Imp J Interdiscip Res 3(4):1253–1256
Das A, Bandyopadhyay S (2010a) Phrase-level polarity identification for Bangla. Int J Comput Linguist Appl 1(1–2):169–182
Das A, Bandyopadhyay S (2010b) Sentiwordnet for Bangla. Knowl Shar Event Task 2:1–9
Das A, Bandyopadhyay S (2010c) Sentiwordnet for Indian languages. In: Asian federation for natural language processing, pp 56–63
Deepamala N, Kumar R (2015) Polarity detection of Kannada documents. In: International advance computing conference. IEEE, pp 764–767
Esuli A, Sebastiani F (2007) Sentiwordnet: a high-coverage lexical resource for opinion mining. In: International conference on language resources and evaluation, pp 1–26
Fondekar A, Pawar JD, Karmali R (2016) Konkani sentiwordnet: resource for sentiment analysis using supervised learning approach. In: Workshop on Indian language data: resources and evaluation (WILDRE3), Portoroz, Slovenia, pp 55–59
Ghosal T, Das SK, Bhattacharjee S (2015) Sentiment analysis on (Bengali horoscope) corpus. In: Annual India conference (INDICON). IEEE, pp 1–6
Govindan R, Haroon RP (2016) A survey on sentiment and emotion classification in Indo-Dravidian languages. Imp J Interdiscip Res 3(1):1040–1042
Gupta CP, Bal BK (2015) Detecting sentiment in Nepali texts: a bootstrap approach for sentiment analysis of texts in the Nepali language. In: International conference on cognitive computing and information processing. IEEE, pp 1–4

Hasan KA, Rahman M et al (2014) Sentiment detection from Bangla text using contextual valency analysis. In: 17th International conference on computer and information technology. IEEE, pp 292–295
Hassan A, Amin MR, Al Azad AK, Mohammed N (2016) Sentiment analysis on Bangla and Romanized Bangla text using deep recurrent models. In: International workshop on computational intelligence. IEEE, pp 51–56
Hegde Y, Padma S (2015) Sentiment analysis for Kannada using mobile product reviews: a case study. In: International on advance computing conference. IEEE, pp 822–827
Hegde Y, Padma S (2017) Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. In: 7th international on advance computing conference. IEEE, pp 777–782
Jayan P, Nair DS, Elizabeth Jisha S (2015) A subjective feature extraction for sentiment analysis in Malayalam language. Int J Eng Sci 14:1–4
Jena MK, Chandra BR (2014) Opinion mining for online Oriya text. Eur J Acad Essays 44–48
Jha V, Manjunath N, Shenoy PD, Venugopal K, Patnaik LM (2015) Homs: Hindi opinion mining system. In: 2nd International conference on recent trends in information systems. IEEE, pp 366–371
Joshi A, Balamurali A, Bhattacharyya P (2010) A fall-back strategy for sentiment analysis in Hindi: a case study. In: Proceedings of the 8th international conference on natural language processing, pp 1–6
Kaur A, Gupta V (2014a) N-gram based approach for opinion mining of Punjabi text. In: International workshop on multi-disciplinary trends in artificial intelligence. Springer, pp 81–88
Kaur A, Gupta V (2014b) Proposed algorithm of sentiment analysis for Punjabi text. J Emerg Technol Web Intell 6(2):180–183
Kaur J, Saini JR (2014) A study and analysis of opinion mining research in Indo-Aryan, Dravidian and Tibeto-Burman language families. Int J Data Min Emerg Technol 4(2):53–60
Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE technical report 2
Kumar A, Kohail S, Ekbal A, Biemann C (2015a) Iit-tuda: system for sentiment analysis in Indian languages using lexical acquisition. In: International conference on mining intelligence and knowledge exploration. Springer, pp 684–693
Kumar KA, Rajasimha N, Reddy M, Rajanarayana A, Nadgir K (2015b) Analysis of users sentiments from Kannada web documents. Procedia Comput Sci 54:247–256
Kumar SS, Premjith B, Kumar MA, Soman K (2015c) Amrita_cen-nlp@ sail2015: sentiment analysis in Indian language using regularized least square approach with randomized feature learning. In: International conference on mining intelligence and knowledge exploration. Springer, pp 671–683
Miranda DT, Mascarenhas M (2016) Kop: an opinion mining system in Konkani. In: International conference on recent trends in electronics. Information and communication technology. IEEE, pp 702–705
Mittal N, Agarwal B, Chouhan G, Bania N, Pareek P (2013) Sentiment analysis of Hindi review based on negation and discourse relation. In: Proceedings of international joint conference on natural language processing, pp 45–50
Mukhtar N, Khan MA (2017) Urdu sentiment analysis using supervised machine learning approach. Int J Pattern Recogn Artif Intell 32(02):1–15
Mukhtar N, Khan MA, Chiragh N (2017) Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis. Cogn Comput 9(4):446–456
Mukhtar N, Khan MA, Chiragh N (2018a) Lexicon based approach outperforms supervised machine learning approach for Urdu sentiment analysis in multiple domains. Telemat Inform 35(8):2173–2183
Mukhtar N, Khan MA, Chiragh N, Nazir S (2018b) Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. Expert Syst 35(6):1–12
Mukku SS, Choudhary N, Mamidi R (2016) Enhanced sentiment classification of Telugu text using ml techniques. In: SAAIP@ 25th international joint conference on artificial intelligence, pp 29–34
Naidu R, Bharti SK, Babu KS, Mohapatra RK (2017) Sentiment analysis using Telugu sentiwordnet. In: International conference on wireless communications signal processing and networking, pp 1–5
Nair DS, Jayan JP, Sherly E et al (2014) Sentima-sentiment extraction for Malayalam. In: International conference on advances in computing, communications and informatics. IEEE, pp 1719–1723
Nair DS, Jayan JP, Rajeev R, Sherly E (2015) Sentiment analysis of Malayalam film review using machine learning techniques. In: International conference on advances in computing, communications and informatics. IEEE, pp 2381–2384
Nivedhitha E, Sanjay S, Anand Kumar M, Soman K (2016) Unsupervised word embedding based polarity detection for Tamil tweets. Int J Comput Technol Appl 9(10):4631–4638
Nongmeikapam K, Khangembam D, Hemkumar W, Khuraijam S, Bandyopadhyay S (2014) Verb based manipuri sentiment analysis. Int J Nat Lang Comput 3(3):113–118
Pandey P, Govilkar S (2015) A framework for sentiment analysis in Hindi using HSWN. Int J Comput Appl 119(19):23–26

Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Found Trends Inf RetR 2(1–2):1–135
Patra BG, Das D, Das A, Prasath R (2015) Shared task on sentiment analysis in Indian languages (sail) tweets-an overview. In: International conference on mining intelligence and knowledge exploration. Springer, pp 650–655
Phani S, IIEST S, Lahiri S, Biswas A (2016) Sentiment analysis of tweets in three Indian languages. In: Proceedings of the 6th workshop on south and southeast Asian natural language processing, vol 1001, pp 93–102
Prasad SS, Kumar J, Prabhakar DK, Pal S (2015) Sentiment classification: an approach for Indian language tweets using decision tree. In: International conference on mining intelligence and knowledge exploration. Springer, pp 656–663
Rani S, Kumar P (2017) A sentiment analysis system to improve teaching and learning. Computer 50(5):36–43
Rani S, Kumar P (2018a) Deep learning based sentiment analysis using convolution neural network. Arab J Sci Eng. https://doi.org/10.1007/s13369-018-3500-z
Rani S, Kumar P (2018b) A sentiment analysis system for social media using machine learning techniques:social enablement. Digit Sch Hum. https://doi.org/10.1093/llc/fqy037
Rehman ZU, Bajwa IS (2016) Lexicon-based sentiment analysis for Urdu language. In: Sixth international conference on innovative computing technology. IEEE, pp 497–501
Rohini V, Thomas M, Latha C (2016) Domain based sentiment analysis in regional language-Kannada using machine learning algorithm. In: International conference on recent trends in electronics, information and communication technology. IEEE, pp 503–507
Sahu S, Behera P, Mohapatra D, Rakesh C (2016a) Information retrieval in web for an Indian language: an Odia language sentimental analysis context. Int J Comput Technol Appl 9(22):249–256
Sahu SK, Behera P, Mohapatra D, Balabantaray RC (2016b) Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative. CSI Trans ICT 4(2–4):111–115
Sarkar K, Chakraborty S (2015) A sentiment analysis system for Indian language tweets. In: International conference on mining intelligence and knowledge exploration. Springer, pp 694–702
Se S, Vinayakumar R, Kumar MA, Soman K (2015) Amrita-cen@ sail2015: Sentiment analysis in Indian languages. In: International conference on mining intelligence and knowledge exploration. Springer, pp 703–710
Se S, Vinayakumar R, Kumar MA, Soman K (2016) Predicting the sentimental reviews in tamil movie using machine learning algorithms. Indian J Sci Technol 9(45):1–5
Seshadri S, Madasamy AK, Padannayil SK (2016) Analyzing sentiment in indian languages micro text using recurrent neural network. IIOAB 7:313–318
Sharma P, Moh TS (2016) Prediction of Indian election using sentiment analysis on Hindi twitter. In: International conference on big data. IEEE, pp 1966–1971
Sharma R, Bhattacharyya P (2014) A sentiment analyzer for Hindi using Hindi Senti Lexicon. In: 11th International conference on natural language processing, pp 1–6
Sharma R, Nigam S, Jain R (2014) Polarity detection movie reviews in Hindi language, pp 1–9. arXiv preprint arXiv:1409.3942
Sharma Y, Mangat V, Kaur M (2015) A practical approach to sentiment analysis of Hindi tweets. In: 1st International conference on next generation computing technologies. IEEE, pp 677–680
Sharmista A, Ramaswami M (2016) Tree based opinion mining in Tamil for product recommendations using R. Int J Comput Intell Inf 6(2):108–116
Syed AZ, Aslam M, Martinez-Enriquez AM (2010) Lexicon based sentiment analysis of Urdu text using SentiUnits. In: Mexican international conference on artificial intelligence. Springer, pp 32–43
Syed AZ, Aslam M, Martinez-Enriquez AM (2011) Sentiment analysis of Urdu language: handling phrase-level negation. In: Mexican international conference on artificial intelligence. Springer, pp 382–393
Syed AZ, Aslam M, Martinez-Enriquez AM (2014) Associating targets with sentiunits: a step forward in sentiment analysis of Urdu text. Artif Intell Rev 41(4):535–561
Thapa LBR, Bal BK (2016) Classifying sentiments in Nepali subjective texts. In: 7th International conference on information, intelligence, systems and applications. IEEE, pp 1–6
Thulasi P, Usha K (2016) Aspect polarity recognition of movie and product reviews in Malayalam. In: International conference on next generation intelligent systems. IEEE, pp 1–5
Venugopalan M, Gupta D (2015) Sentiment classification for Hindi tweets in a constrained environment augmented using tweet specific features. In: International conference on mining intelligence and knowledge exploration. Springer, pp 664–670