Artif Intell Rev (2014) 41:535–561
DOI 10.1007/s10462-012-9322-6

Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text

Afraz Z. Syed · Muhammad Aslam ·
Ana Maria Martinez-Enriquez

Published online: 13 March 2012

Abstract This paper presents, a grammatically motivated, sentiment classification model, applied on a morphologically rich language: Urdu. The morphological complexity and flexibility in grammatical rules of this language require an improved or altogether different approach. We emphasize on the identification of the SentiUnits, rather than, the subjective words in the given text. SentiUnits are the sentiment carrier expressions, which reveal the inherent sentiments of the sentence for a specific target. The targets are the noun phrases for which an opinion is made. The system extracts SentiUnits and the target expressions through the shallow parsing based chunking. The dependency parsing algorithm creates associations between these extracted expressions. For our system, we develop sentiment-annotated lexicon of Urdu words. Each entry of the lexicon is marked with its orientation (positive or negative) and the intensity (force of orientation) score. For the evaluation of the system, two corpora of reviews, from the domains of movies and electronic appliances are collected. The results of the experimentation show that, we achieve the state of the art performance in the sentiment analysis of the Urdu text.

Keywords Natural language processing · Sentiment analysis · Opinion mining · Shallow parsing · Dependency parsing

A. Z. Syed (✉) · M. Aslam
University of Engineering and Technology, Lahore, Pakistan
e-mail: afrazsyed@uet.edu.pk

M. Aslam
e-mail: maslam@uet.edu.pk

A. M. Martinez-Enriquez
Department of CS, CINVESTAV-IPN, Mexico, D.F., Mexico
e-mail: ammartin@cinvestav.mx1

1 Introduction

The Web 2.0 has emerged, as a platform for the dynamic information exchange and the personal view propagation. Now, more and more people around the globe express their feelings through blogs, give voice to the governmental and political affairs through news reviews, and record their likes and dislikes in the form of product reviews. This proliferation of the information has affected the lives of the internet users both positively as well as negatively. On one side, the people use internet forums, blogs, consumer reports, product reviews, and different type of discussion groups for taking everyday decisions. But, on the other hand, the negative aspect of this sharing opinion may not be ignored.

According to Glaser et al. (2002) the extremist groups use Internet to endorse hatred and aggression. Internet has turn into a ubiquitous, anonymous, economical, and rapid way of communication for such groups (Crilley 2001). Therefore, the analysis of user generated web content is not only useful for commercial purposes, but also, its need for the discouragement of such misinformation is more immediate, particularly, in the main languages of the world.

Consequently, the research on opinion mining and sentiment analysis on some Indo-European languages, like, English, is flourishing and have a number of successful contributions (Turney 2002; Pang et al. 2002; Riloff et al. 2003; Riloff and Wiebe 2003; Tan et al. 2009 and Bloom and Argamon 2010).

It is not yet decided whether and how equivalent success could be attained for Morphologically Rich Languages (MRLs) (Abdul-Mageed and Korayem 2010). MRL takes into consideration the syntactic units. The relationships are expressed at word-level, i.e., the structure of the word is complex and morphological operations like inflection and derivation are more frequent (Tsarfaty et al. 2010). Due to this word level complexity, the MRL becomes more challenging for the computational linguistic (CL) applications. Urdu is a worth mentioning case in this point. Besides Urdu is a major language with about 100 million speakers, there is also a great potential for performing the sentiment analysis on.

Urdu language is morphologically rich, its constituent words and phrases tend to be more complex, due to the recurrent derivations and inflections. Besides, the morphological complexity, the variability in the grammar rules and vocabulary in the Urdu text is usual and is considered acceptable. Urdu is influenced by many other languages (e.g., Hindi, Persian, Arabic, Sanskrit and English) not only in vocabulary but also in morphology and grammar. The loanwords from a particular language follow their own grammar rules. Hence, Urdu language has distinctiveness in features and in linguistic aspects. Moreover, Urdu is altogether different from the well recognized languages in the field of sentiment analysis and other CL applications.

Our approach is grammatically motivated, incorporating a sentiment-annotated lexicon for the identification of the sentiment carrier expressions in a sentence. The expressions are labeled as SentiUnits (Syed et al. 2010), which reveal the inherent sentiments of the sentence for a specific target. For instance, consider two sentences, “300 is a *terrific* movie.” and “300 is a *not that much remarkable,* as was expected.” In both statements, the italic words are labeled as the SentiUnits. The over all sentence subjectivity is based on these expressions, whereas the other terms are considered neutral. Therefore, the subjective polarity of a sentence is computed by the polarities of its constituent SentiUnits.

The sentiment-annotated lexicon based classifier in our previous effort (Syed et al. 2010) focuses on (*a*) the extraction of the SentiUnits, (*b*) the computation of the polarity scores of the sentences according to the extracted SentiUnits, and (*c*) the classification of the review according to these polarity scores. This approach is good for handling the sentences with single targets. In other words, it can only handle simple opinions in which all the opinionated expressions are associated with one object or target. Presence of multiple targets, as in the comparative sentences, where two different targets are compared, may lead to a misclassification error, e.g., It is hard to rank 300 among the *outstanding* movies like Brave Heart, or Ben-Hur". In this case, the analyzer may misclassify the comment. As, the expression, "*outstanding*" is positive and is by default associated to the movie "300", which is presented for review. This is because the analyzer is not establishing an expression to target link. The positive expression "*outstanding*" should be linked with the movies "Brave Heart and Ben-Hur", instead of the reviewed movie "300".

To handle this kind of misclassification in complex sentences like comparatives, in this paper we extend the model and divide a single opinionated sentence into three units: a source of appraisal, a SentiUnit (the appraisal expression), and finally a target of this appraisal (Bloom and Argamon 2010; Whitelaw et al. 2005).

To minimize misclassification rate, our present approach emphasizes on the precise identification of SentiUnits as well as their associated targets. To associates each SentiUnit with its respective target, a new module called the *ASSOCIATOR is included*. The *ASSOCIATOR* module uses the dependency parsing based algorithm. The *EXTRACTOR* module uses shallow parsing based chunking to extract the SentiUnits (Syed et al. 2010).

The performance of the system was evaluated on the corpus of reviews about movies and electronic appliances. We have used four performance metrics: precision, recall, and F-measure in addition to accuracy. In comparison, with our previous version, the results are radically improved with 82.5% of accuracy, particularly for sentences with multiple targets.

Let us denote the review under consideration as R in Urdu text. R is single sentence based or it contains multiple sentences, among which some are subjective sentences in the set $S_s = \{S_{s1}, S_{s2}, S_{s3,\ldots.} S_{sk}\}$ and others are objective $S_o = \{S_{o1}, S_{o2}, S_{o3,\ldots.} S_{ol}\}$, such that,

$$R = \left\{S_{s1}, S_{s2}, S_{s3,\ldots.} S_{sk}\right\} U \left\{S_{o1}, S_{o2}, S_{o3,\ldots} S_{ol.}\right\},$$

where,

$$k = 1, 2, 3, \ldots n;\ l = 1, 2, 3, \ldots m;\ n \text{ and } m \text{ are finite numbers.}$$

A polarity value P_s is allocated for each sentence through PREPROCESSOR, EXTRACTOR, and ASSOCIATOR modules of the system (Sect. 3). The final polarity of the review P_R is calculated as a sum of all sentence polarities by CLASSIFIER module:

$$P_R = \sum P_{si,} \quad \text{where } i = 1, 2, 3, \ldots N;\ N \text{ is a finite number.}$$

The goal is to develop and integrate a whole sentiment analysis model for Urdu text. To achieve this goal the following issues were tackled:

1. Unluckily, no annotated Urdu lexicon available exists, thus we extended and modified the lexicon version developed in Syed et al. (2010) to this purpose. The sentiment-annotated Urdu lexicon includes information about subjectivity of entries in addition to orthographic, phonological, syntactic, and morphological aspects.
2. Due to the morphological complexity of Urdu, algorithms used on languages like English (Pang and Lee 2008; Wiebe et al. 2004; Bloom and Argamon 2010), Chinese (Jang and Shin 2010), Arabic (Abbasi et al. 2008) cannot be applied directly to process Urdu. Thus, the processing and the classification of Urdu text in accordance with the inherent sentiments were designed and constructed (Sect. 4).

Finally, in order to validate our approach, for the experimentation, we have used sentiment annotated lexicon of Urdu words and two corpora of reviews about movies and electronic appliances as test-beds. The performances of the extended model and the previous version are given and the results are found significantly improved (Sect. 5).

2 Challenge: distinctive features of the Urdu language

The distinctiveness of a language is recognized by its vocabulary, orthography, and morphology. Here we present these features of the Urdu language.

2.1 Vocabulary

In addition to Arabic, Persian, and Turkish influences, Urdu kept on including the vocabulary from English, Sanskrit, and Hindi. Hence, the absorption power of Urdu is very exceptional, enhancing the magnificence of the language. Some examples of Urdu words taken from other languages are shown in Table 1, along with the use in the sentences.

2.2 Orthography

Urdu uses Persio-Arabic script, which is cursive and context-sensitive with respect to the shapes of the alphabets. It means that the "حروف" (*haroof*, alphabets) have multiple glyphs and shapes, which are categorized as joiners and non-joiners. The joiner alphabets join together into units, called the *ligatures* (Durrani and Hussain 2010). One word can have either single or multiple ligatures. During writing, all characters join together until a non-joiner appears. A new ligature starts after the non-joiner. The process is repeated until the word ends.

According to the position (initial, medial, final) in the ligature the Urdu character exhibits multiple shapes or it remains unconnected. For example, the alphabet "ج" (*jeem*) can be joined in initial position as "جا", in medial position as "بجا" and at final position as "حج", see Table 2.

Table 1 Examples of Urdu words from multiple languages

Language	Borrowed words	Example of Urdu sentences
English	ٹیلی فون (*telephone*, Telephone)	ٹیلی فون خراب ہے (*telephone khrab hay*, Telephone is out of order)
Persian	فردوس (*firdos,* heaven)	سوات فردوس نظیر ہے (*sawat firdos nazeer hay*, Sawat is like heaven)
Sanskrit	آشا (*aasha,* wish)	میری آشا پوری ہوگئ (*meri aasha puri ho gayee*, My wish came true)
Turkish	خاتون (*khatoon*, lady)	وہ ایک نفیس خاتون ہیں (*woh aik nafees khatoon hain*, She is a fine lady)
Arabic	جنت (*janat,* heaven)	گھر جنت ہے (*ghar janat hay*, Home is heaven)

Table 2 Different shapes of a single alphabet "ج" (jeem)

Remark	Shape adjustment
Joined in the initial position	ا + ج → جا
Joined in the medial position	ا + ج + ب → بجا
Joined at the final position	ج + ح → حج
In a word with a non-joiner	ج + آ → آج

Table 3 Examples of morphological processes in Urdu

Operation	Word	Modified form
Inflection	پھول (*phool*, flower)	پھولوں (*phool-on*, flowers)
Derivation	ممکن (*mumkin*, possible)	ناممکن (*na -mumkin*, impossible)
Compounding	جان (*jaan*, soul), دل (*dil*, heart)	دل و جان (*dil-o -jaan*, heart and soul)
Partial Reduplication	رات (*raat*, night)	راتوں رات (*raat-on -raat*, in a night)
Compound verbs	مار (*maar*, beat), ڈالو (*dalo*, put)	مار ڈالو (*maar dalo*, kill)

Due to this context sensitive orthography and difference in the behaviors of joiners and non-joiners, the word boundary identification becomes a major task. The space is not always an indicator of the word boundary.

2.3 Morphology

Urdu language comes in the category of MRL like Arabic, Turkish, Finnish, Persian, and Chinese's. Word-structure is complex, because syntactic units and relations are expressed at word-level. Morphological operations like, inflection and derivation are more frequent. Some morphological processes for are discussed below and shown in Table 3:

(i) *Inflection and derivation.* Inflectional operation deals with the variety of forms of the same words. The changes indicate grammatical features, e.g., "جانا" (*jana*, to go) from "جا" (*ja*, go). The diversity of these inflections implies much complexity. For instance, in Urdu language, the Arabic loan words are made plural according to Arabic grammar, whereas, the Persian loan words follow the Persian grammar, e.g., the plural of "لفظ" (*lafz*, word) is "الفاظ" (*alfaaz*, words) and "پودا" (*poda*, plant) is "پودے" (*poday*, plants). To make plural word, both are differently inflected. By contrast, plural in English, according to the predefined grammatical rules, is obtained by only adding *s*, *es* or *ies*. Exceptions exist but they are rare. Derivational operations deal with the production of new words with different meanings. New words are produced by adding affixes. Often the produced words have a changed part of speech, e.g., "خوش" (*khush*, happy) and "خوش بخت" (*khushbakht*, lucky).

(ii) *Compounding.* The compounding process generates new words by combination of two already existing words M and N. Some examples of compound words in Urdu are:

Table 4 Examples of affixes, case markers and postpositions

1. Morphemes	پودا → پودے plural postfix ے (*ay*) is applied
2. Words or lexical units	نا (*na*), نی (*ni*), نے (*nay*), سے (*say*)
2.1. Case marker	سے (*say*), کو (*ko*)
2.1.1. Core case markers	میں نے کہا (*mein nay kaha*, I said)
2.1.2. Oblique case marker	باہر نکالنا (*bahir nikal- na*, put out)
2.2. Pure postpositions	کمرے میں جا (*kamray mein ja*, go to the room)
2.3. Possession or genitive markers	آپ کا نام (*aap ka naam*, your name)

- *MN* formation: *M* and *N* are independent in meaning and syntax, they are only written together, making a new word. For example, $M =$ "موم" (*mom*, wax), N = "بتی" (*bati*, light), producing the word $MN =$ "موم بتی" (*mombati*, candle).
- $M - O - N$ formation: *M* and *N* are independent words, but are related in meaning or context. Their syntax remains the same with an additional alphabet "و" (O) that means "and". For example, $M =$ "ملک" (*mulk*, country), $N =$ "ملت" (*milat*, nation), make the compound word, $M - O - N =$ "ملک و ملت" (*mulk-o-milat*, country and nation).

(iii) *Reduplication.* Both full and partial reduplication of words is very common in Urdu. For example, the full reduplication of the word "کبھی" (*kabhi*, sometime), result into "کبھی کبھی" (*kabhi kabhi*, infrequently).

(iv) *Compound Verbs or Verb Phrases.* In Urdu root verbs and intensifying verbs combine together to form compound verbs (Schmidt 1999). For example, the root verb "پکار" (*pukar*, call) and intensifying verb "لو" (*lo*, take) make a compound verb "پکار لو" (*pukarlo*, call (right away)). This compound verb has the same meaning as the root verb but exhibit more strength.

2.4 Independent case marking

Urdu text contains two types of affixes: (*a*) morphemes and (*b*) words or lexical units. Morphemes are lexically attached with the nouns through morphological operations. For example, to make plural "پودے" (*poday*, plants) of the word "پودا" (*poda*, plant) plural postfix "ے" (*ay*) is applied as shown in Table 4.

While the words or lexical units are independent units. These are further categorized as case markers, pure postpositions and possession or genitive markers. The case markers are further divided into core case markers and oblique case markers. They mark grammatical function to the marked words and are generally, morphologically attached with the words at the lexical level. But, in Urdu, they are syntactically attached and lexically independent. The property of free word-order in Urdu text is due to the case markers, which can identify constituents in multiple ways (Rizvi and Hussain 2005).

As an example of core case markers consider the sentence, "میں نے کہا" (*mein nay kaha*, I said), in which the case marker "نے" (*nay*) is used. Similarly, in the sentence "آپ کا نام" (*aap ka naam*, your name), the possession marker "کا" (*ka*) is used. Table 4 gives some more examples.

Thus, Urdu is a challenging language with diverse vocabulary, cursive and context sensitive orthography, and complex morphology. Therefore, Urdu language processing has its own requirements related to the task of sentiment analysis and should be studied as an independent problem domain. Hence, our sentiment classification model presented in Sect. 3, handles these issues effectively. For example the problem of word segmentation is handled through the PREPROCESSOR module and is presented in Sect. 3.1.

3 The sentiment classification model

Our sentiment classification model is able to handle MRL like Urdu. Our model is grammatically motivated and employs a sentiment-annotated lexicon based classification approach for the identification of the sentiment carrier expressions in a sentence, called the SentiUnits. The sentence subjectivity is based on these expressions and all other terms are considered neutral. The subjective polarity of a sentence is computed by the polarities of its constituent SentiUnits. A single opinionated sentence is logically partitioned into: a source of appraisal, a SentiUnit (the appraisal expression), and finally a target of this appraisal (Bloom and Argamon 2010; Whitelaw et al. 2005).

Firstly, the SentiUnits and the targets are extracted, and then the targets are associated with the respective SentiUnits.

The sentiment analysis breaks up into four modules (see Fig. 1). The *PREPROCESSOR* module identifies the word boundaries and segments the sentence into the meaningful words or lexical units. The out put of *PREPROCESSOR* is the input of the *EXTRACTOR* module. *EXTRACTOR* extracts the sentiment expressions and the noun phrases, as SentiUnits and Targets, respectively. The *ASSOCIATOR* module is responsible for linking the candidate targets to each extracted SentiUnit. Finally, the *CLASSIFIER module* identifies polarities of each SentiUnit for each sentence and calculates the overall sentiment of each review as a sum of sentence polarities.

3.1 PREPROCESSOR

In general, for NLP applications, the preprocessing phase removes punctuation marks, omitting unnecessary symbols and striping of HTML tags. Besides *PREPROCESSOR* module for our application has to handle the *diacritics omission* and *word boundary* identification issues, which are specific to Urdu language:

3.1.1 Diacritic omission

Similar to the other Arabic script based languages (Persian, Turkish, Sindhi, and Punjabi), Urdu script consists of two classes of symbols: letters and diacritics. Just like the letters, the diacritics are also useful for readability and understanding of the script. They not only represent the vowels, but also affect the meanings of the words. However in writings, these symbols are optional and this is observed that some authors use some diacritics regularly and others are totally ignored. Even the over use of a particular kind is very common.

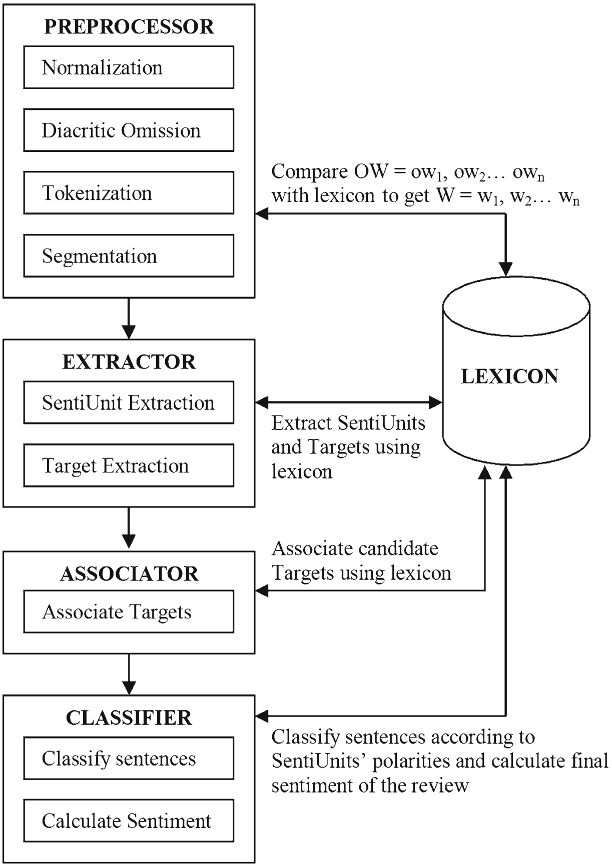
PREPROCESSOR
Normalization
Diacritic Omission
Tokenization
Segmentation
Compare $OW = ow_1, ow_2 \ldots ow_n$ with lexicon to get $W = w_1, w_2 \ldots w_n$
LEXICON
EXTRACTOR
SentiUnit Extraction
Target Extraction
Extract SentiUnits and Targets using lexicon
ASSOCIATOR
Associate Targets
Associate candidate Targets using lexicon
CLASSIFIER
Classify sentences
Calculate Sentiment
Classify sentences according to SentiUnits' polarities and calculate final sentiment of the review

Fig. 1 Modules of the system model

Hence, their use is highly author dependent. This under and over use and some times absence of diacritics adds to the morphological as well as lexical ambiguity of the language. For example, the task of POS tagging of the diacritic bearing words can generate incorrect results due to ambiguous meaning. This critical problem is considered as unresolved in linguistics research. In consequence, a regular practice, so the diacritics are removed as a part of preprocessing phase (Durrani and Hussain 2010).

3.1.2 Word boundary identification

In almost all NLP applications, word segmentation or word boundary identification through tokenization is the foremost obligatory task. Tokenization is easy to implement for languages whose word boundaries are identified through punctuation marks or white spaces, e.g., Spanish, English, and French (Lehal 2010).

Thus, the input is considered as a sequence of letters, which determine a sequence of the words, i.e., $< w_1, w_2, w_3 \ldots w_i > \rightarrow < l_1, l_2, l_3 \ldots l_j >$. Each sentence is segmented into lexical words based on *word boundaries*. But, this process becomes complicated, if white spaces or other word delimiters are rarely or never used as word boundaries.

As we already mentioned in Sect. 2, Urdu orthography is context sensitive. The "حروف" (*haroof*, alphabets) are divided in two categories as joiners and non joiners. The joiners take multiple glyphs and shapes according to the context, which cause word boundaries identification issues. In (Durrani and Hussain 2010) the word segmentation of Urdu text is divided into two sub problems: space insertion and *space deletion*.

Space-insertion. Many words in Urdu are made by more than one ligature (usually two). Semantically and syntactically these ligatures are part of a single word. If the last letter of the first ligature in a word is a joiner then it tends to join with the first letter of the second ligature. To avoid this joining, the writer inserts a space. This causes space insertion errors, e.g., "خوش باش" (*khush bash*, happy), is a single word with two ligatures, L_1= "خوش" and L_2 = "باش". The last letter of L_1 "ش" is a joiner which tends to join with first letter in L_2 "ب" to avoid this joining a space is inserted while typing the word. On omitting this space we get "خوشباش", whish is not a correct word, thus, the space cannot be avoided.

Space-omission. There are many words which end with non-joiner letters. As the non-joiner letters keep a constant shape so usually the writers do not insert spaces while writing the next word to identify word boundary. This does not affect the readability of the words but for computational tasks the boundary identification becomes an issue as both words are written in continuation without space. For example, the phrase, "شیراوربکری" (*shair aur bakri*, lion and goat) is written without space, and "شیر اور بکری" (*shair aur bakri*, lion and goat) is written with spaces. We resolved this issue by including into the phrase the symbol "|" to indicate the word boundaries "شیر |اور |بکری" (*shair aur bakri*, lion and goat).

The word segmentation for the Urdu text is considered by most of the researches as a major task, since includes morphological analyzer, POS tagger, and translators. A few contributions dealt with this issue as an independent task (Durrani and Hussain 2010; Lehal 2010, 2009). Particularly, Durrani and Hussain (2010) presents a detailed survey for the identification of the inherent causes and proposes a word segmentation model.

Thus, PREPROCESSOR in our application performs four steps (see Fig. 2). First, the normalization operation is performed, which removes symbols and tags. Then the diacritic omission is performed to avoid ambiguity. Thirdly, the sentence is tokenized as a sequence of orthographic words: $OW = \{ow_1, ow_2 \ldots ow_n\}$ where the words $ow_1, ow_2, \ldots$ are not grammatical or meaning full words but these are only orthographically separated each other. This sequence becomes the input to the Segmentation module. The Segmentation result is the sequence of meaning full and grammatically correct words ready for further processing.

3.2 EXTRACTOR

The *EXTRACTOR* module identifies and obtains the SentiUnits and the targets. Two subtasks are performed (see Fig. 4): (a) Extracting SentiUnits with Adjectives as head words; (b) Extracting targets with Nouns as head words.

3.2.1 Extracting SentiUnits with adjectives as head words

SentiUnits can be defined as the core grammatical structures, expressing the opinion or the sentiment carrier expressions in a sentence (Syed et al. 2010). For understanding the structure of the SentiUnits, consider the examples shown in Table 5.

In Table 5, the underlined expressions are responsible for subjectivity orientation. All other words are neutral and have no effect into the classification. On a closer look at these

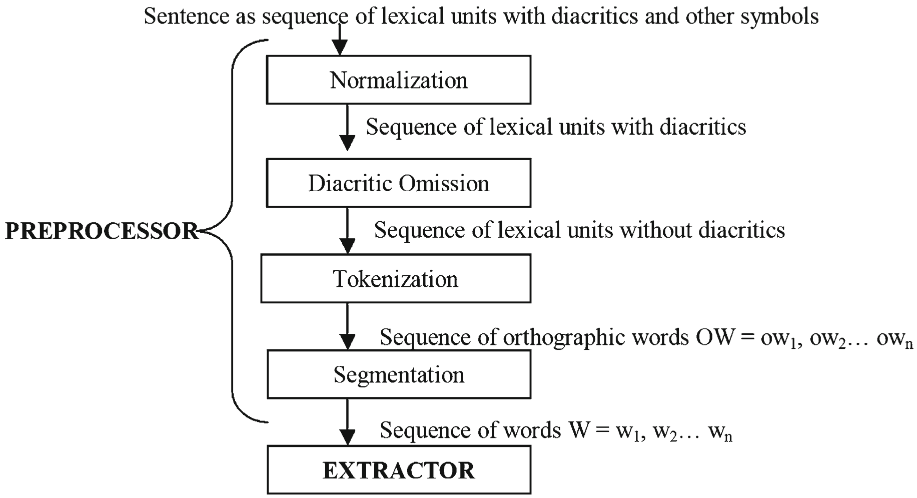
Sentence as sequence of lexical units with diacritics and other symbols
Normalization
Sequence of lexical units with diacritics
Diacritic Omission
Sequence of lexical units without diacritics
PREPROCESSOR
Tokenization
Sequence of orthographic words OW = ow1, ow2… own
Segmentation
Sequence of words W = w1, w2… wn
EXTRACTOR

Fig. 2 Preprocessing of the input sentence by the PREPROCESSOR module

Table 5 Examples of opinionated sentences from Urdu with different SentiUnits

This is a fine book.	*Yeh aik umdah kitab hay*	1 یہ ایک عمدہ کتاب ہے .
This is a fine and informative book.	*Yeh umdah aur malumati kitab hay*	2 یہ عمدہ اور معلوماتی کتاب ہے .
This is the finest book.	*Yeh sab se umdah kitab hay*	3 یہ سب سے عمدہ کتاب ہے .
This book is not very bad.	*Yeh kitab itni buri naheen*	4 یہ کتاب اتنی بری نہیں .

examples, we can observe that the SentiUnits are made of adjectives (as head words). These can be single word/adjective based like sentence 1, or multiple words based like sentences 2, 3, and 4. Also, the sentence 1, 2, and 3 have adjectives with positive orientation, but sentence 4 contains a negative word but due to the use of negation it becomes positive. In this case, negation acts as a polarity shifters. Moreover, the intensity of the expressions is determined by modifiers which can be absolute, comparative, or superlative just like English text. Sentence 3 represents the example of the superlative degree of the appraisal.

Hence, these expressions can be distinguished by six attributes: adjectives as the head words, their modifiers, and their orientation towards positive or negative, the intensity of this orientation, a polarity mark assigned to each word to show the intensity value and finally the negation (Syed et al. 2010). Here, we discuss these attributes in detail;

Adjectives in Urdu can be divided into two types (Schmidt 1999) (see Fig. 3). First, those describing quantity and quality, e.g. "کم" (*kam*, less), "بدترین" (*budtareen*, worst), "زیادہ" (*ziyada*, more). The second type of adjectives distinguishes one from others, e.g. "حسین" (*haseen*, pretty), "فطین" (*fateen*, intelligent).

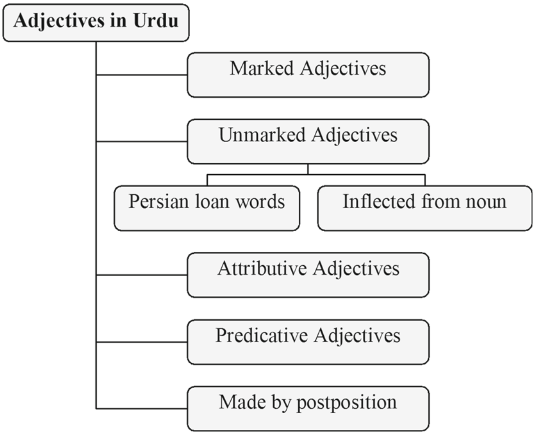

Fig. 3 Types of adjectives in Urdu

Further adjectives are categorized as *marked* and *unmarked* (Schmidt 1999). *Marked* are those which can be inflected for number and gender, e.g., (a) "اچھا کام" (*acha kaam*, good work), (b) "اچھے کام" (*achay kaam*, good works) and (c) "اچھی بات" (*achi aadat*, good habit). In (a), (b), and (c) "اچھا" (*acha*, good) is inflected for masculine, plural and feminine, respectively. *Unmarked* are usually Persian loan words, e.g., "تازہ" (*tazah*, fresh) and the adjectives inflected from nouns, e.g., "دفتری" (*daftary*, official) inflected from "دفتر" (*daftar*, office). *Attributive* adjectives are very frequent and they precede the noun they qualify, e.g., the adjective "مزیدار" (*mazedaar*, tasty) precede the noun "مزہ" (*maza*, taste). Arabic and Persian loan adjectives are used predicatively, e.g., "معلوم ہونا" (*maloom hona*, to be Known) (see Fig. 3). These adjectives appear in the form of phrases. The postposition "سے" (*say*), "سی" (*si*), "سا" (*sa*) and "والا" (*wala*), "والی" (*wali*), "والے" (*walay*) are frequently used with noun to make adjectives, e.g., "پھول سی" (*phool si*, like flower) from "پھول" (*phool*, flower). (see Fig. 3).

The *modifiers* intensify the orientation of an adjective. These can be absolute, comparative or superlative. The modifiers made by postpositions are very frequent in Urdu writing. For example, the absolute adjective "مہنگا" (*mehnga*, expensive) is modified by the postposition "سے" to make it comparative; "اس سے مہنگا" (*is say mehnga*, more expansive).

The postposition "سب سے" makes a superlative expression; "سب سے مہنگا" (*sab say mehnga*, most expansive). Some Persian loan words are also commonly used in inflected forms. For example, "کم" (*kam*, less) is absolute and is inflected to make comparative "کمتر" (*kamtar*, lesser) and superlative "کمترین" (*kamtareen*, least) expressions. Detailed examples of modifiers are given in Table 6.

The *negations* words like "no, not, do not, don't, can't" can altogether alter the sense of a sentence so are very important to tackle. They are polarity shifters, e.g. "موسم اچھا ہے" (*mosam acha hay*, the weather is pleasant.) becomes "موسم اچھانہیں ہے" (*mosam acha naheen hay*, the weather is not pleasant.). Different approaches are used to handle negations, e.g., Hu and Liu (2004) processes negations as a part of post processing and associates

Table 6 Adjective modifiers

Modifier	Made by postpositions	Persian loan words
Absolute	مہنگا (*mehnga*, expensive)	کم (*kam*, less)
Comparative		
(a) سے	اس سے مہنگا (*is say mehnga*, more expansive)	کمتر→ کم + تر
(b) سے زیادہ	اس سے زیادہ مہنگا (*is say ziyadah mehnga*)	*kam* + *tar*→ (*kamtar,* lesser)
Superlative		
(a) سب سے	سب سے مہنگا (*sab say mehnga*, most expansive)	کمترین→ کم + ترین
(b) سب میں	سب میں مہنگا (*sab main mehnga*)	*Kam* + *tareen* → (*kamtareen,* least)
(c) سب سے زیادہ	سب سےزیارہ مہنگا (*sab say ziyadah mehnga*)	

negating words with the subjective components of the sentence using co-location. This technique works in sentences like "*I don't like*", "*This is not good*" is not effective in the sentence "*No doubt it is amazing*" (Pang and Lee 2008). Whitelaw et al. (2005) considers negations as part of appraisal expressions annotated with attitudes. In this way, the negations became independent of location. Another approach is the use of POS tagged corpus (Na et al. 2004).

3.2.2 Extracting targets with Nouns as head words

EXTRACTOR identifies the targets through shallow parsing based chunking. These targets are the non-overlapping noun phrases "اسمی ترکیب" (*ismi tarkeeb*) present in the text. Noun phrases are the units of one or more words in a link with noun as head word and all other words as dependents. Urdu noun phrases exhibit variations in structure and complexity level. Even a noun phrase can include other phrases as its components, e.g., adjectival and genitive phrases etc. In addition to internal complexity of the noun phrase its position in the sentence is not always the same. This is due to the free word order property of Urdu text (Rizvi and Hussain 2005). Hence, the chunker for Urdu noun phrases must be capable of handling both aspects simultaneously.

3.2.3 Working of the module

Shallow parsing based text chunking is used by this module. This method identifies the beginnings and ends of grammatical phrases without parsing the full phrase structure. Hence, the *EXTRACTOR* shallow parses each sentence in the given review to find adjective or noun phrases and then work out for attributes (modifiers, orientation, intensity) modeling the behavior of modifiers and negations within the phrase.

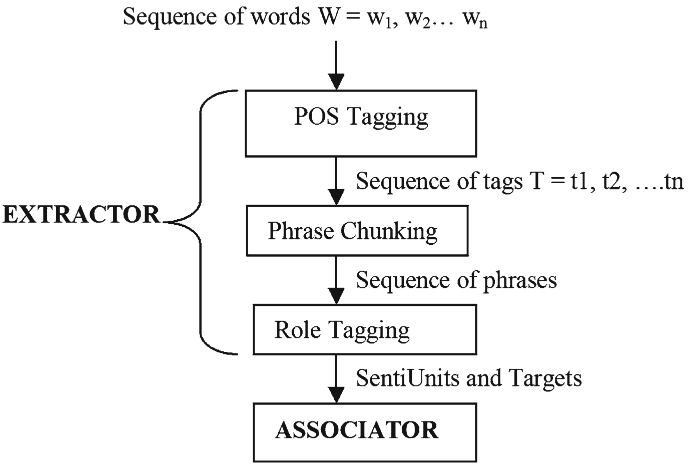

Fig. 4 Processing of the input sentence by *EXTRACTOR* module

For extracting SentiUnits, the parser starts with a lexicon of nominal and adjectival head words, which define initial values for orientation whether positive or negative. In addition to positive or negative orientation head words exhibit the intensity of orientation. It searches for occurrences of these head words in the sentence, and upon finding them it moves rightward to attach modifiers because the modifiers appear in the right side of the adjectives in Urdu. Now, the parser searches for the polarity shifters or negations and finally distinguishes the whole subjective expression. Likewise the parser identifies candidate targets with the help of lexicon. It finds the entire target groups matching words specified in the lexicon. These steps are shown in Fig. 4. The following examples describe the execution of the model given in Fig. 4.

Example 1 “ارتضی کا روبوٹ بڑا شاندار ہے”

(*Irtaza ka robot bara shaandaar hay,* Irtaza’s robot is very fabulous.)

In this sentence, both the SentiUnit and the target are complex, since they are composed of more than one word. The SentiUnit “بڑا شاندار” (*bara shaandaar,* very fabulous) is made by an adjective head word and a positive modifiers. The target of the comment “ارتضی کا روبوٹ” (*Irtaza ka robot,* Irtaza’s robot) is based on three words: two nouns with a possession marker in between, as shown in Table 7.

Example 2 “ارتضی اورفاطمہ کا کمرہ ہوادارنہیں”

(*Irtaza aur fatima ka kamrah hawadar naheen,* Irtaza and fatima’s room is not airy)

Again, both the SentiUnit and the target are complex. The SentiUnit “ہوادارنہیں” (*hawadar naheen,* not airy) contains an adjective head and a negation word. The target of the comment is even more complex, i.e., “ارتضی اورفاطمہ کا کمرہ” (*Irtaza aur Fatima ka kamrah,* Irtaza and fatima’s room) since it is made by five words: three nouns, a possession marker and a conjunction. The sentence parse in given in Table 8.

Table 7 Parsing of example 1 into targets and SentiUnits

Remark	Parse	
Sentence with complex SentiUnit (SU) and target (NP)	[N PM N] [ADJ ADJ] AUX → NP SU AUX	ارتضی کا روبوٹ بڑا شاندار ہے
Noun phrase (NP) with possession marker (PM)	N PM N→ NP (Target)	ارتضی کا روبوٹ
SentiUnit made by two adjectives (ADJ)	ADJ ADJ → SU (SentiUnit)	بڑا شاندار

Table 8 Parsing of example 2 into targets and SentiUnits

Remark	Parse	
Sentence with complex SentiUnit and target	[N CJC N PM N] [ADJ NEG] → NP SU	ارتضی اورفاطمہ کا کمرہ ہوادارنہیں
Noun phrase with conjunction (CJC) and possession marker (PM)	N CJC N PM N →NP (Target)	ارتضی اورفاطمہ کا کمرہ
SentiUnit with negation (NEG)	ADJ NEG → SU (SentiUnit)	ہوادارنہیں

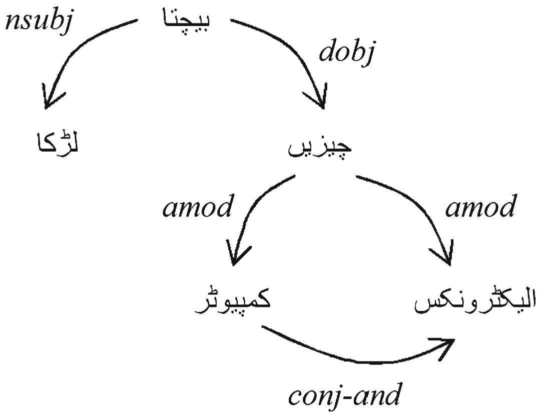

Fig. 5 The dependency parsing of the given sentence

3.3 ASSOCIATOR

The extracted SentiUnits and targets are associated each other through *ASSOCIATOR*. We apply dependency parsing for this purpose. Figure 5 shows the dependency parsing of the sentence; "لڑکا کمپیوٹر اور الیکٹرونکس کی چیزیں بیچتا ہے" (*larka computer aur electronics kee cheezain baichta hay.* The boy sells computer and electronic products).

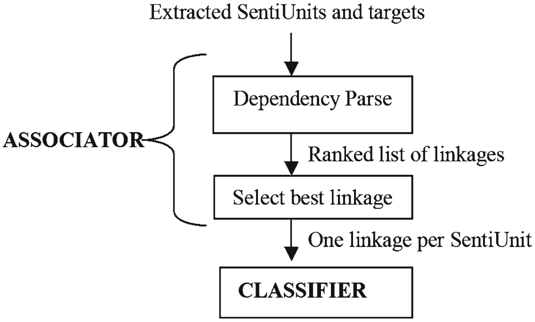

Fig. 6 Linking SentiUnits with candidate targets by *ASSOCIATOR* module

3.3.1 Working of ASSOCIATOR

First the nominal group that is the lexical representation of the target is identified and then the values of the attributes describing that target are computed. ASSOCIATOR finds the target phrase by following the paths through a dependency parse of the sentence. The result of the dependency parse is a ranked list of paths or linkage specifications. These specifications are ranked to specify the order in which the links should be traversed. For each SentiUnit, the system looks for the paths through the dependency tree which annotate any word in the SentiUnit to the next or final expected word according to the specification of that particular link. With the identification of a word in the proper syntactic place, the shallow parsing is applied moving rightward to find a noun phrase that ends in the identified word. These steps are shown in Fig. 6.

The steps performed by *ASSOCIATOR* are:

Input: Shallow parsed sentence with extracted SentiUnits and targets.
Processing: Apply dependency parse and then,

1. Search all the linkages such that;
   a. The linkage is in the linkage specifications
   b. The linkage connects to a chunked SentiUnit
   c. The linkage need not connect to chunked target
2. For each chunked SentiUnit;
   a. If any linkage to the chunked target exists then,
   b. Remove unconnected linkages
3. Select the linkage according to priority of linkage specifications.

Output: One linkage per SentiUnit.

For example, take the linkage specification shown below:

$$\text{Target} \xrightarrow{nsubj} \text{a} \xleftarrow{dobj} \text{b} \xleftarrow{amod} \text{SentiUnit}$$

We apply it to the sentence "بانگ درا ایک اچھی کتاب ہے" (*bang-e-dara aik achi kitab hay.* Baang-e-Dara is a good book.), the chunker finds "اچھی" *as* a sentiment expression. The ASSOCIATOR module then searches for the target noun phrase, which is "بانگ درا", the name of the book (see Fig. 7).

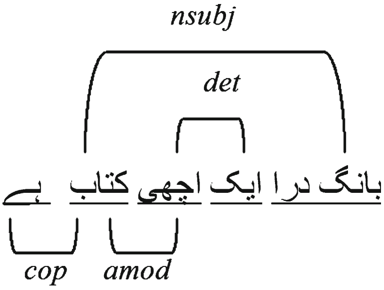

Fig. 7 Linking the sentiment expressions with candidate targets

3.4 CLASSIFIER

The *CLASSIFIER* module starts from calculating the intensity of orientation of the SentiUnits by comparing each tagged word with the polarity values assigned in the lexicon entries. For example, the expression "بہت اچھی کتاب" (*bohat achi kitab*, very good book) is more intense than "اچھی کتاب" (*achi kitab*, good book) due to the modifier "بہت", (*bohat*, very) and both are positive expressions. In this expression, the SentiUnit "بہت اچھی" (*bohat achi*, very good) is associated with the target "کتاب" (*kitab*, book). *CLASSIFIER* look for other associations identified by ASSOCIATOR, then it calculates the polarity value for each association for a particular target, e.g., "کتاب" (*kitab*, book) in this case. If "بہت اچھی" (*bohat achi*, very good) is the only expression in the sentence showing sentiments about the target then the sentence polarity is equal to the polarity of this expression otherwise other possible expressions are also evaluated. The calculation of polarity is summation of either positive or negative expressions with positive or negative values respectively.

3.4.1 Working of the CLASSIFIER

According to the problem statement, the given review, R may be a single sentence based or it may contain multiple sentences, among which, some may be subjective sentences in the set $S_s = \{S_{s1}, S_{s2,} S_{s3,....} S_{sk}\}$ and others are objective $S_o = \{S_{o1}, S_{o2,} S_{o3,....} S_{ol}\}$, such that,

$R = \{S_{s1}, S_{s2,} S_{s3,....} S_{sk}\} U \{S_{o1}, S_{o2,} S_{o3,...} S_{ol.}\}$, where, $k = 1, 2, 3, \ldots n$; $l = 1, 2, 3, \ldots, m$; where n and m are finite.

The final polarity of the review P_R is calculated as a sum of all sentence polarities computed by the *CLASSIFIER* module. If P_{si} represent the sentence polarities of i sentences then,

$$P_R = \sum P_{si,} \text{ where } i = 1, 2, 3, \ldots N; \; N \text{ is a finite number.}$$

Hence, the *CLASSIFIER* module is composed by two steps:

Step1: Compute sentence polarity
Input: Dependency parsed sentence with SentiUnits to targets associations.
Process: Start with any one SentiUnit of a particular target

a. COMPARE each word in the SentiUnit with the lexicon to find its orientation and polarity value;
b. COMPUTE SentiUnit polarity by adding polarities of the words according to the intensity values;

```
    c. LOOK FOR another SentiUnit for the same target;
    d. Sentence polarity = SUMMATION of all SentiUnits' polarities for a particular target;

Step2: Compute total polarity of review

    a. REPEAT step 1 for all sentences
    b. ADD all polarity values to calculate $P_R$
    c. COMPARE with threshold

      Case a: If $P_R >$ threshold, then $R$ is positive.
      Case b: If $P_R <$ threshold, then $R$ as negative

Output: Classification of positive or negative review.
```

We have presented our approach and the platform composed by *PREPROCESSOR, EXTRACTOR, ASSOCIATOR,* and *CLASSIFIER* modules (see Fig. 1). Next the evaluation of the model through experimentation is presented.

4 Evaluation

Due to analytical evaluation of the sentiment classifier requires a formal specification of the problem with respect to how correctness and completeness are defined; the evaluation of the sentiment classifiers is conducted experimentally. Besides the practical effectiveness and performance of the classifier does not emphasize on.

On the other hand, the experimental evaluation of a classifier usually measures its effectiveness in terms of its ability to take the accurate classification decisions.

Hence, we have performed a series of four experiments in two sets. The results are given in Sect. 4.3. The lexicon and corpora are discussed in Sects. 4.1 and 4.2, respectively.

4.1 Lexicon

Lexicon is a main requirement for most of the NLP applications. For sentiment analysis, lexicon becomes more complex, because, it contains sentiments annotated to all entries in addition to their grammatical, morphological, and phonological information. Besides, Urdu is a resource poor language and the construction of such a lexicon become even more laborious task. This task includes multiple aspects, for instance, identification of the sentiment oriented expressions in Urdu language, syntactic structures, and morphological rules, e.g. inflection or derivation. The grammatical rules like; use of modifiers. Likewise, identification of the relationships between the lexicon entries, e.g. synonyms, antonyms and cross references, polarities, possible modifiers and the intensities of these modifiers. The lexicon originally developed in Syed et al. (2010) was with the perspective of SentiUnits only. With the extension of the model we extend the lexicon as well. We have extended it to incorporate nominal appraisal head words, and modifiers.

Currently version of the lexicon contains 1,368 adjectives, which are marked according to the orientation and the intensity. There are 67 modifiers, including both comparative and superlative intensity levels. The nominal head words are selected according to the domains of the movies and the electronic appliances, which are 1,920 in number. A summary of the existing version of the lexicon is given in Table 9.

4.2 Corpus

Due to the deficiency of publicly accessible corpus of the Urdu language based reviews, we collect two corpora of reviews to evaluate the efficacy of the employed model. The first corpus $C1$ is the collection of 700 movie reviews, among which 385 are positive and 315 are negative. The average document length in this corpus is 264 words. For obtaining variant reviews, 40 different movies with different popularity scores (already known) and categories (comedy, drama, historical etc) are given for review.

The second test-bed is a corpus of reviews of the electronic appliances $C2$. This corpus comprises a total of 650 reviews with 322 positive and 328 negative. The base collection has the reviews for three types: refrigerators (237), air-conditioners (250), and televisions (163). The average review length is 196 words. For achieving diversity, 9 different brands of the electronic appliances are given for review.

For both corpora, the reviews within the threshold boundary or with neutral scores are removed. Hence, the data set contains either positive or negative reviews as shown in Table 10.

4.3 Results

For evaluating the effectiveness and efficiency of a text classifier only using the accuracy as the performance metric is not sufficient. We use other three metrics; called the precision P, recall R and F-measure F in addition to accuracy A.These metrics can provide much greater insight into the performance features of a classifier. For a sentiment classifier the accuracy A can be defined as the measure of how close the document classification suggested by the classifier is, to the actual sentiments present in the review. The precision P measures the exactness of a classifier. A higher P means less false positive and vice versa. Whereas, the recall R measures the sensitivity or completeness of the classifier. Higher R means less false negative and vice versa. In terms of true positive t_p, false positive f_p, true negative t_n and false negative. P and R can be defined as:

$$P = t_p / \left(t_{p+} f_p\right)$$
$$R = t_p / \left(t_{p+} f_n\right)$$

F-measure is produced by combining Precision and Recall, which is the weighted harmonic mean of both values, as defined below:

$$F = 2PR/(P + R)$$

Table 9 Summary of lexicon entries

Modifiers	Adjectival head words	Nominal head words
67	1,368	1,920

Table 10 Corpora for evaluation

Domains	Total number	Average length (words)	Orientation	Number
Movies C1	700	264	*Positive*	385
			Negative	315
Electronic appliances C2	650	196	*Positive*	322
			Negative	328

Table 11 Experimental results in terms of P, R, F and A for model A

Orientation	Corpora	Precision	Recall	F-measure	Accuracy (%)
Positive	$C1$	0.737	0.681	70.8	74
	$C2$	0.795	0.737	76.5	79
Negative	$C1$	0.698	0.654	67.5	66
	$C2$	0.785	0.767	77.6	77

Table 12 Comparison of accuracy from both corpora C1 and C2 for model A

Corpora	Orientation	Accuracy (%)	Variation (%)	Corpora accuracy (%)	Total accuracy (%)
$C1$	*Pos*	74	8	70	74
	Neg	66			
$C2$	*Pos*	79	2	78	
	Neg	77			

Table 13 Experimental results in terms of P, R, F and A for model B

Orientation	Corpora	Precision	Recall	F-measure	Accuracy (%)
Positive	$C1$	0.822	0.795	80.8	80
	$C2$	0.897	0.877	88.7	88
Negative	$C1$	0.795	0.777	78.6	77
	$C2$	0.865	0.832	84.8	84

A series of four experiments in two sets with two models of the system have been performed. The *model A* is the former version of the system and the *model B* is the current version in which the ASSOCIATOR module is attached. By using this testing, the efficacy and usability of the extended version are easily compared. Both models are applied on both corpora $C1$ and $C2$ separately.

Model A: Tables 11 and 12 show the results of the experiments performed by m*odel A* on both corpora $C1$ and $C2$. Table 11 shows the detailed results with P, R, F and A values separately computed for positive as well as negative reviews.

Table 12 shows a comparative summary of the results from both corpora. The accuracy of $C1$ is 70% and variation in positive and negative reviews is 8%. Whereas the accuracy of $C2$ is 78% and variation in positive and negative reviews is 2%. The total accuracy of *model A* is 74%.

Model B: For the next two experiments we include ASSOCIATOR module and tested both corpora. The results are shown in Tables 13 and 14. Table 13 shows the experimental results in terms of *P, R, F,* and *A* for *model B* applied on $C1$ and $C2$ for positive and negative reviews separately.

Results from Table 13 are compared and summarized in Table 14. The accuracy of $C1$ improves to 78.5%, and the variation in positive and negative reviews decreases to 3%. Likewise, the accuracy of $C2$ increases to 86.5% and the variation in the accuracy of positive and negative reviews also increases to 3%. The total accuracy of *model B* is 82.5%.

Table 14 Comparison of accuracy from both corpora C1 and C2 for model B

Corpora	Orientation	Accuracy (%)	Variation (%)	Corpora accuracy (%)	Total accuracy (%)
$C1$	*Pos*	80	3	78.5	82.5
	Neg	77			
$C2$	*Pos*	88	3	86.5	
	Neg	85			

From the above results it is clear that the classification accuracy is highly domain specific. The reviews in $C1$ are more challenging to classify as compared to those of electronic appliances in $C2$. The reason is that these reviews contain more allegory which results into more divergence, not only syntactic or semantic structure, but also in appraisal type. Discussion about the movie plot and its characters weather good or evil is very frequent phenomenon. This discussion results into a number of appraisal targets which further can lead to the selection of the wrong linkage. On the other hand all positive or negative comments about the parts of an electronic appliance are indirectly related to the same target.

Moreover, the classification accuracy also depends upon the orientation of the review. From results, it is also perceptible that negative reviews are more prone to be misclassified than the positive ones.

5 Sentimental analysis: state of the art

The sentiment analysis has influenced from different domains like information retrieval, data mining, computational linguistics. The foregoing efforts have covered a broad range of tasks at different granularity levels: polarity classification at document level (Pang et al. 2002); opinion identification at sentences level classification (Pang and Lee 2004); emphasize on phrases (Turney 2002); opinion source assignment at phrase level (Breck et al. 2007; Choi and Cardie 2008).

To present a precise a literature survey, we focus of two aspects of sentiment classification: (a) features of the given text on the basis of which, the classification algorithms are developed; (b) techniques or methods, used for the implementation of the algorithms.

5.1 Features

Researchers have focused on a number of features of the given text for achieving better classification results. The features are encoded into vectors for the proper application of machine learning algorithms (Pang and Lee 2008). Thus, feature selection is a critical task and can affect the results to a great extend. Syntactic, semantic, linking based, term based, topic oriented and part of speech based features are frequently used in literature. We discuss four categories: Part of speech (POS) based, term based, syntactic, and topic oriented.

The POS based information, particularly, adjectives, can help in sentiment analysis. That is why the earliest work in this domain uses adjectives as subjectivity indicators (Hatzivassiloglou and McKeown 1997). After that, Hatzivassiloglou and Wiebe (2000), Mullen and Collier (2004) and Whitelaw et al. (2005) handle adjectives using multiple techniques. Turney (2002) argues that, proverbs are also carriers of sentiments in a sentence and should be considered in combination with adjectives. The sentences are divided into pre-structured grammatical patterns, which include adjectives and adverbs as the core words. Riloff et al. (2003) attempts a relatively new idea and proposes the analysis of nouns in the text. It emphasizes on the concept of subjective nouns and computes the orientation for the phrases in the sentence which contained them.

Many works are available in which term based features are considered. For example, the position of the term in a sentence is put forward as a feature by Kim and Hovy (2006). This work locates the specific terms, and then, according to their position, it computes subjectivity orientation. Another work, Wiebe et al. (2004) applies the concept of *hapax legomenon* for feature selection, which means, a word occurring only once in a given corpus. It proposes that such words tend to be more subjective than the others. In addition to this feature, it uses a relatively complex syntactic feature, i.e., collocations of the words in a sentence. If some words or terms co-occur more frequently than usual, then, these are considered as collocations. According to Yang et al. (2006) the terms which are rare and are not entered in a prefixing dictionary tend to be more subjective, because the reviewers use them to emphasis their opinion.

Pang et al. (2002) states better performance, using "presence of term" a*s a* binary-valued feature vector, whose entries merely specify, whether a term occurs (0, 1) or not. But, in a term frequency feature vector entry values increase with the occurrence frequency of the corresponding term (Abdul-Mageed and Korayem 2010). Bigrams and trigrams are used by Dave et al. (2003). Kennedy and Inkpen (2006) and Snyder and Barzilay (2007) consider contrastive distance between terms as an automatically computed feature. Whitelaw et al. (2005) uses the concept of appraisal theory and extracts appraisal expressions with the help of sentiment lexicon. Mullen and Collier (2004) observes that, the sentences which contain a reference to the topic can be considered more important. For this purpose, it specifies words and word phrases which, can be extracted as indicators of the reference. The discussed features and contributions with examples are summarized in Table 15.

5.2 Techniques

There are techniques used for sentiment analysis like unsupervised bootstrapping, sentiment lexicon, and support vector machines (see Table 16). In unsupervised bootstrap approach, a primary or initial classifier is applied on the text to generate labeled data as the output. After that, a supervised learning algorithm may be applied on this data. The initial classifier can have various implementation possibilities, according to the language complexity and depth of the required analysis. An example of such an initial high-precision classifier to learn extraction patterns for subjective terms is proposed by Riloff and Wiebe (2003). Kaji and Kitsuregawa (2007) uses this method for the automatic construction of HTML documents based corpus in which, the polarity labels are assigned to the entries.

Hatzivassiloglou and Wiebe (2000), Turney (2002), Yu and Hatzivassiloglou (2003), Riloff et al. (2003) and Higashinaka et al. (2006) employ sentiment-annotated lexicon induction technique. As a first step, an unsupervised approach is applied for the generation of a sentiment-annotated lexicon. Then using this as a resource, the given text is classified as positive or negative. This technique is further discussed in Sect. 3.

Hu and Liu (2004) and Andreevskaia and Bergler (2006) use Preston WordNet for extraction of sentiment tags. There is also a trend in research community to extend existing lexicons, e.g. SentiWordNet is an extension of the WordNet.

Table 15 Features used and their respective contributions

Type	Focused features	Contributions
Term based	Term presence and position	Pang et al. (2002)
	Bigrams and trigrams	Dave et al. (2003)
	Hapax legomena	Wiebe et al. (2004)
	Rare terms for emphasis	Yang et al. (2006)
	Tem position	Kim and Hovy (2006)
	Term frequency	Abdul-Mageed and Korayem (2010)
	Contrastive distance in terms	Kennedy and Inkpen (2006)
		Snyder and Barzilay (2007)
Syntax based	Collocations	Riloff and Wiebe (2003)
		Wiebe et al. (2004)
	Appraisal expressions	Whitelaw et al. (2005)
	Valance shifters	Kennedy and Inkpen (2006)
	Noun adjective dependency	Bloom and Argamon (2010)
POS based	Adjectives	Hatzivassiloglou and McKeown (1997)
		Hatzivassiloglou and Wiebe (2000)
		Mullen and Collier (2004)
		Whitelaw et al. (2005)
	Adjective and adverb	Turney (2002)
	Subjective noun	Riloff et al. (2003)
Topic	Reference to the topic	Mullen and Collier (2004)

Table 16 Techniques used by different contributions

Technique used	Contributions
Unsupervised bootstrapping	Riloff and Wiebe (2003)
	Kaji and Kitsuregawa (2007)
Sentiment annotated lexicon	Hatzivassiloglou and Wiebe (2000)
	Turney (2002)
	Yu and Hatzivassiloglou (2003)
	Riloff et al. (2003)
	Higashinaka et al. (2006)
Support vector machines (SVM)	Pang et al. (2002)
	Dave et al. (2003)
	Pang and Lee (2004)
	Kennedy and Inkpen (2006)
WordNet based	Hu and Liu (2004)
	Andreevskaia and Bergler (2006)

5.3 Sentiment-annotated-lexicon construction

As we are using the lexicon based approach for the development of the sentiment analyzer so we discuss here some contributions from this aspect the research. Lexicon construction with an apposite coverage is a challenging task. From definition of grammar rules to their appropriate implementation, it requires much expertise and proficiency about the target language as well as the computer algorithms. For the task of sentiment analysis the entries of these lexicons are annotated with the orientation scores in addition to their morphological, grammatical and phonological information. This sentiment annotation task can either be done manually with the help of the agreement of judges who can decide about the orientation scores of the given words. Or, it can be done automatically, using computer algorithms like machine learning approaches etc. The manual annotation, provides higher accuracy but is more time consuming and lengthy.

The languages, which are more popular on the internet, have rich and easily available electronic-linguistic-resources. For example, English language, for which almost all types of corpora are available from almost all domains, i.e., from product reviews to news discussions. That is why the sentiment analysis research community has moved to the algorithms and approaches which can help in the generation of the automatic lexicons as an alternative of manual annotation and tagging. For example, Annet and Kondrak (2008), Higashinaka et al. (2006), Andreevskaia and Bergler (2006), Hu and Liu (2004), Yu and Hatzivassiloglou (2003), Riloff et al. (2003), Turney (2002) and Hatzivassiloglou and Wiebe (2000). These methods are fast and can rapidly develop domain dependent lexicons.

Going back to the history of sentiment annotated lexicon construction, General Inquirer (Stone et al. 1966) is a popular recourse for sentiment analysis of English language and is manually compiled. A pioneering attempt in automatic acquisition of sentiment annotated-lexicon is Hatzivassiloglou and McKeown (1997). This work develops a sentiment-annotated lexicon with an emphasis on adjectives. It presents a scheme based on the conjunctions between the adjectives in a large corpus. They apply shallow parsing algorithm and developed a log-linear statistical model. This model predicts same orientation between any two adjectives. After that automatic acquisition of the polarity values of words and phrases itself appeared as an active line of research. Diverse techniques have been proposed and implemented for learning the word polarities. These include corpus-based approaches like Hatzivassiloglou and McKeown (1997), statistical approaches to measures of the word association etc as proposed in Turney and Littman (2003) and using lexical relationships (Kamps et al. 2004).

Some other efforts have tried to use or extend the existing lexicons, e.g. the extension of WordNet is SentiWordNet. In SentiWordNet the polarity marks are annotated with the existing structure of the gloss. Andreevskaia and Bergler (2006) and Hu and Liu (2004) utilize WordNet or its extensions for the sentiment analysis. Moreover, Hatzivassiloglou and Wiebe (2000), Turney (2002), Yu and Hatzivassiloglou (2003), Riloff et al. (2003) and Higashinaka et al. (2006) have tried to develop algorithms and techniques for automatic lexicon construction using unsupervised learning methods.

Most of these efforts use pre-developed linguistic recourses like corpuses for the development and extraction of required lexicons. But, Urdu is a recourse poor language and hence the task of lexicon construction becomes more difficult and time consuming. To our knowledge no such lexicon exists for Urdu text. However, there are a very few efforts who have tried to construct corpuses and simple lexicons for other NLP applications.

The preliminary work is presented for the EMILLE (Enabling Minority Language Engineering) project in the form of a multi-lingual corpus for the South Asian languages. A parallel corpus for Hindi, Urdu, English, Bengali, Punjabi and Gujarati languages contains about 200,000 words (Baker et al. 2003). Their independent corpus of Urdu text has 1,640,000 words annotated with POS tags (Hardie 2003).

Another effort is presented in Ijaz and Hussain (2007). They use corpus to automatically develop Urdu lexicon. Their corpus is based on cleaned text from news websites, containing about 18 million words. The work Muaz et al. (2009), gives brief analysis of parts of speech of Urdu language and develops a POS tagged corpora, whereas, another effort (Mukund et al. 2010) generates semantic role labeled corpus for Urdu text using cross lingual projections. Humayoun et al. (2007) presents the extraction and development of the automatic extraction of Urdu lexicon using corpus.

5.4 Generalization

The generalization of sentiment analysis solutions, among multiple domains is still an open issue. The term *domain adaptation* is coined by the SA community (Tan et al. 2009) to refer to the development of a generalized solution which can be applied on all the potential target domains. Most of the contributions for opinion mining are highly domain specific (Pang and Lee 2004). Tan et al. (2009) handles the domain adaptation issue using frequency co-occurring entropy (FCE) method. **It** emphasizes on a smooth transformation from a domain d_1 to another domain d_2 through a set of generic features F, representing d_1 and d_2. **It** evaluates the model for six domains and finally concludes that FCE is not the best option. Another feature related to multiple domains is their complexity level. Sentiment analysis of reviews related to products and movies is considered as the easiest in literature (Pang and Lee 2004) and these reviews serve as a test bed for most of the approaches. On the contrary, political speeches and discussions are perhaps the most complex to handle. Bansal et al. (2008) pinpoints an issue and evaluates whether the speech is in favor or opposition.

5.5 Sentiment analysis of MRL

There are some worth mentioning contributions for handling sentiment analysis in MRL. For example, Abdul-Mageed and Korayem (2010) and Abbasi et al. (2008) for Arabic, and Jang and Shin (2010) for Chinese language, etc. The work presented in Abdul-Mageed and Korayem (2010) is for sentiment analysis of the Arabic text. In this work, the main focus is on the Arabic text related issues for the development of a practical analyzer with acceptable performance. It analyzes news text by automatic classification at the sentence level. It applies a support vector machines classifier. Another related work is Abbasi et al. (2008). It performs sentiment analysis of Arabic and English web forums. Its emphasis is on the extremist opinion propagation. For handling Arabic language's characteristics, it proposes specific feature extraction components. It develops Entropy Weighted Genetic Algorithm (EWGA), a hybridized genetic algorithm that incorporates the information gain heuristic for feature selection, i.e., stylistic and syntactic features. This algorithm improves the system performance by selecting better key features.

5.6 Sentiment analysis and Urdu language processing

The field of Urdu language processing is not yet well explored and well established, this is due to the distinctive linguistic features of the language as discussed in the Sect. 2. Particularly, in the sentiment analysis of the Urdu text (Syed et al. 2010) emphasizes on the extraction of the adjective based expressions, from the given text. Apart from sentiment analysis there are of course some considerable contributions, which pave the ways to accomplish this task.

Most of the contributions made by Center for Research in Urdu Language Processing are of major significance for Urdu computational linguistics (Durrani and Hussain 2010; Ijaz and Hussain 2007; Muaz et al. 2009). Durrani and Hussain (2010) highlights the issue of Urdu word boundary identification or word segmentation. Moreover, Ijaz and Hussain (2007) presents core concerns in the development of Urdu lexicon. Another contribution given in

Mukund et al. (2010) uses cross-lingual projection for the generation of labeled corpus for Urdu. It labels the corpus with semantic roles and achieved good results. Another effort is a case-marking model of Urdu-Hindi languages by using semantic information (Rizvi and Hussain 2005). Urdu and Hindi morphology is implemented in Xerox finite state technology based on ASCII transliteration and Functional Morphology (FM) is used for implementing Urdu morphology (Humayoun et al. 2007).

We have precisely discussed, the various features of the given text, like, POS based, term based, syntactic, and topic oriented. Similarly, we have presented a comprehensive overview of classification techniques used by different contributions, e.g., unsupervised bootstrapping, sentiment lexicon and support vector machines.

Since it is obvious from the above discussion that the POS based features particularly adjectival phrases are considered more frequently and successfully that is why we emphasize in this research on the adjectival phrases and we call them SentiUnits. Moreover, we use the sentiment lexicon based technique for classification of reviews.

6 Conclusions

The principal contribution of our research is the study of Urdu language as an independent problem domain, since Urdu differs from English in script, morphology, and grammar. Despite of similarities with Arabic and Persian script, and with Hindi morphology, Urdu language has its own requirements as far as computational linguistics is concerned.

Our sentiment analysis framework employs a grammatical model based approach. This approach focuses on the sentence grammatical structures, besides to the morphological structure of the words. Primarily, two types of grammatical structures (adjective phrases as SentiUnits and nominal phrases as their targets) are extracted and then linked. The extraction and linking is achieved by implementing two parsing methods: shallow and dependency parsing, respectively. As a result of this effort, 82.5% of accuracy was obtained. Our infrastructure is easy to extend by including more features that characterize different sentence constituents and this becomes our future endeavor. Another potential future effort is the domain adaptation for a generalized MRL sentiment analyzer.

References

Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst, pp 1–34

Abdul-Mageed M, Korayem M (2010) Automatic identification of subjectivity in morphologically rich languages: the case of Arabic. In: Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (WASSA), Lisbon, pp 2–6

Andreevskaia A, Bergler S (2006) Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses. In Proceedings of the 11th conference of the European chapter of the association for computational linguistics, EACL-2006, Trent, pp 209–216

Annet M, Kondrak G (2008) A comparison of sentiment analysis techniques: polarizing movie blogs. In: Proceedings of Canadian AI, pp 25–35

Baker P, Hardie A, McEnery T, Jayaram BD (2003) Corpus data for South Asian language processing. In: Proceedings of the EACL workshop on South Asian languages, Budapest

Bansal M, Cardie C, Lee L (2008) The power of negative thinking: exploring label disagreement in the min cut classification framework, Manchester. In: Proceedings of COLING, pp 13–16

Bloom K, Argamon S (2010) Unsupervised extraction of appraisal expressions. In: Proceedings of Canadian AI, Ottawa, pp 290–294

Breck E, Choi Y, Cardie C (2007) Identifying expressions of opinion in context. In: Proceedings of IJCAI'07. Menlo Park, CA, pp 2683–2688

Choi Y, Cardie C (2008) Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the conference on empirical methods in natural language processing, Honolulu, HI, pp 793–801

Crilley K (2001) Information warfare: new battle fields, terrorists, propaganda, and the Internet. ASLIB Proc 53(7):250–264

Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the twelfth international world wide web conference (WWW 2003), Budapest, pp 519–528

Durrani N, Hussain S (2010) Urdu word segmentation. In: Proceedings of 11th annual conference of the North American chapter of the association for computational linguistics, Los Angeles

Glaser J, Dixit J, Green DP (2002) Studying hate crime with the Internet: What makes racists advocate racial violence? J Soc Issues 58(1):177–193

Hardie A (2003) Developing a tagset for automated part-of-speech tagging in Urdu. In: Proceedings of the conference of the corpus linguistics, Lancaster

Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of ACL'97. Stroudsburg, PA, pp 174–181

Hatzivassiloglou V, Wiebe JM (2000) Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th international conference on computational linguistics, New Brunswick, NJ

Higashinaka R, Prasad R, Walker MA (2006) Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the ACL, Sydney, pp 265–272

Hu M, Liu B (2004) Mining and summarizing customer reviews. In *Proceedings of SIGKDD'04*, pp 168–177

Humayoun M, Hammarström H, Ranta A (2007) Urdu morphology, orthography and lexicon extraction. In: Proceedings of the 2nd *workshop on computational approaches to Arabic script-based languages. Stanford, USA*, pp 59–66

Ijaz M, Hussain S (2007) Corpus based Urdu lexicon development. In: Proceedings of the conference on language technology, University of Peshawar, Pakistan

Jang H, Shin H (2010) Language-specific sentiment analysis in morphologically rich languages. In: Proceedings of the COLING, Poster Volume, Beijing, pp 498–506

Kaji N, Kitsuregawa M (2007) Building lexicon for sentiment analysis from massive collection of html documents. In: Proceedings of EMNLP'07, pp 1075–1083

Kamps J, Marx M, Mokken RJ, de Rijke M (2004) Using Wordnet to measure semantic orientation of adjectives. In Proceedings of LREC'04, pp 1115–1118

Kennedy A, Inkpen D (2006) Sentiment classification of movie and product reviews using contextual valence shifters. Comput Intell 22(2):110–125

Kim S-M, Hovy E (2006) Automatic identification of pro and con reasons in online reviews. In: Proceedings of the COLING, Sydney, pp 483–490

Lehal GS (2009) A two stage word segmentation system for handling space insertion problem in Urdu script. In: Proceedings of world academy of science, engineering and technology, Bangkok, pp 321–324

Lehal GS (2010) A word segmentation system for handling space omission problem in Urdu script. In: Proceedings of the 1st workshop on South and Southeast Asian natural language processing (WSSANLP), the 23rd international conference on computational linguistics, COLING, Beijing, pp 43–50

Muaz A, Ali A, Hussain S (2009) Analysis and development of Urdu POS tagged corpora. In: Proceedings of the 7th workshop on Asian language resources, ACL-IJCNLP, Suntec, Singapore, pp 24–31

Mukund S, Ghosh D, Srihari RK (2010) Using cross-lingual projections to generate semantic role labeled corpus for Urdu—a resource poor language. In: Proceeding of the 23rd international conference on computational linguistics COLING, Beijing, pp 797–805

Mullen T, Collier N (2004) Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the conference on empirical methods in natural language processing, Barcelona, pp 412–418

Na J-C, Sui H, Khoo C, Chan S, Zhou Y (2004) Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In: Proceedings of conference of the international society of knowledge organization (ISKO), pp 49–54

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd meeting of the association for computational linguistics, Barcelona, pp 271–278

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retrieval 2(1–2):1–135

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in NLP, Philadelphia, PA, pp 79–86

Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Sapporo, pp 25–32

Riloff E, Wiebe J, Wilson T (2003) Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the 7th conference on natural language learning, Edmonton, pp 25–32

Rizvi SMJ, Hussain M (2005) Modeling case marking systems of Urdu-Hindi languages by using semantic information. In: Proceedings of natural language processing and knowledge engineering, pp 85–90

Schmidt RL (1999) Urdu: an essential grammar. Routledge Publishing, New York

Snyder B, Barzilay R (2007) Multiple aspect ranking using the Good Grief algorithm. In: Proceedings of the joint human language technology/North American chapter of the ACL conference, Rochester, NY, pp 300–307

Stone PJ, Dunphy DC, Smith MS, Ogilvie DM (1966) The general inquirer: a computer approach to content analysis. MIT Press, Cambridge

Syed AZ, Muhammad A, Martínez-Enríquez AM (2010) Lexicon based sentiment analysis of Urdu text using SentiUnits. In: Proceedings of the 9th Mexican international conference of artificial intelligence, Pachuca, Mexico, pp 32–43

Tan S, Cheng X, Wang Y, Xu H (2009) Adapting Naive Bayes to domain adaptation for sentiment analysis. In: Proceedings of the 31st *European conference on IR research on advances in information retrieval*, pp 337–349

Tsarfaty R, Seddah D, Goldberg Y, Kübler S, Candito M, Foster J, Versley Y, Rehbein I, Tounsi L (2010) Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In: Proceedings of the NAACL HLT 2010 first workshop on statistical parsing of morphologically-rich languages, Los Angeles, pp 1–12

Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of 40th meeting of the association for computational linguistics, Philadelphia, PA, pp 417–424

Turney P, Littman M (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Trans Inf Syst 21(4):315–346

Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for sentiment analysis. In: Proceedings of ACM SIGIR conference on information and knowledge management (CIKM 2005), Bremen, pp 625–631

Wiebe J, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Comput Linguist 30(3):277–308

Yang K, Yu N, Valerio A, Zhang H (2006) WIDIT in TREC 2006 Blog Track. In: Proceedings of Text REtrieval conference—TREC

Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of EMNLP’03, pp 129–136