Adjectival Phrases as the Sentiment Carriers in the Urdu Text

Afraz Z. Syed[1], Aslam Muhammad[2], Martinez-Enriquez A. M.[3]

[1,2]Department of CS & E, U. E. T. Lahore, Pakistan
[3]Department of CS, CINVESTAV-IPN, D.F. Mexico
([1]afrazsyed@uet.edu.pk, [2]maslam@uet.edu.pk, [3]ammrtin@cinvestav.mx)

Abstract. In this paper we present a comprehensive overview of the structures of the adjectival phrases in the Urdu language with respect to the task of sentiment analysis. Urdu is a widely spoken but one of the least explored languages by the computational linguistics community. After a detailed analysis of adjectival phrases in Urdu text we conclude that this language is orthographically, morphologically and grammatically different from other well established languages, like English and hence, it requires updated or different approaches and algorithms for the task of sentiment analysis. We present our approach in which the adjectival phrases are combined with polarity shifters, and conjunctions to make sentiment expressions in the opinionated sentences. We label these sentiment expressions as the SentiUnits. We apply shallow parsing based chunking to extract the SentiUnits. The overall polarity of a sentence in a given review can be determined by computing the polarity of these expressions. Adjectives are the head words, which appear with modifiers and postpositions. The experimentation based evaluation of the model with a sentiment-annotated lexicon of Urdu words and two corpuses of reviews as test-beds, shows encouraging achievement in terms of sentimental analysis and accuracy.
[Afraz Z. Syed, Aslam Muhammad, Martinez-Enriquez A. M. Adjectival Phrases as the Sentiment Carriers in the Urdu Text. Journal of American Science 2011;7(3):644-652]. (ISSN: 1545-1003). http://www.americanscience.org.

Keywords: Natural language processing, computational linguistics, sentiment analysis, opinion mining, shallow parsing, Urdu text processing.

1. INTRODUCTION

An adjective is a fundamental part of speech (POS) that expresses an attribute of a noun (place, thing or, person). Generally in the sentence structure adjectives appear in two ways, whether they are directly linked with the noun within the noun phrase or they associate with the noun through some other part of speech, e.g., verb. In both cases they describe the characteristic features of the noun they qualify. This point suggests that any opinion, sentiment, or judgment about a noun can be determined by analyzing its adjectives. Due to this characteristic the first effort for the automatic sentiment analysis (SA) of the English text employ adjectives as the main feature of the given text (Hatzivassiloglou & McKeown, 1993). Therefore, in sentiment analysis community, adjectives remain center of attention (Turney, 2002), (Riloff et al., 2003), (Riloff & Wiebe, 2003) and (Bloom & Argamon, 2010).

As with all parts of speech, in every language the use, type, and structure of the adjectives differ. Urdu is morphologically rich and hence its adjectives and adjectival phrases tend to be more complex, due to the frequent inflections and derivations. In addition to the morphological complexity the variability in vocabulary and grammar rules in Urdu text is regular and is considered normal. This is due to the fact that this language is strongly influenced by many other languages like, Persian, Arabic, Sanskrit and English. For example, the adjective "تازہ" (tazah, fresh) remain unmarked because it is Persian loan word and follow Persian grammar, whereas, most of the Sanskrit based adjectives show inflection to agree with the noun they qualify. For example, the demonstrative adjective "جیسا" (*jaisa*, such as), becomes " جیسے" (*jaisee*, such as) and " جیسی" (*jaisay*, such as) for gender and number, respectively. Moreover, the use of post positions as independent lexemes involves more specific patterns and rules.

These aspects suggest that Urdu have distinct characteristics and features. Particularly, it is far more different from the well established languages in the field of sentiment analysis or other NLP applications. The SA research community requires a complete understanding of the computational as well as linguistic aspects of the language. We therefore, present in this paper a comprehensive overview of the structures of the adjectival phrases in the Urdu text with respect to the task of sentiment analysis. For Urdu based NLP research this is the very first effort. So far, syntactic and morphological aspects of the language are considered related to verbs, nouns and other parts of speech (Muaz et al., 2009), (Riaz, 2007), and (Durrani & Hussain, 2010). There is no endeavor about analyzing the sentiments in the given text. Also,

we find no contribution which investigates Urdu adjectival phrases independently.

The given analysis covers almost all the aspects of adjectives and adjectival phrases. We describe their morphological structures, as marked and unmarked through the types of the agreement with the noun they qualify. This agreement is more frequent for gender, number, and case. Also, we discuss their structure when used with a sequence of nouns and for the formations of reduplications. Moreover, we define and illustrate with examples different adjective classes, i.e., descriptive, predicative, attributive, possessive, demonstrative, and reflexive possessive. For each class we describe the morphological structure of the adjectives and their inflected forms. We take most commonly used adjectives as examples and clearly describe their modifications.

Given an inclusive overview of the adjectival phrases, we present our proposal for the sentiment analysis of Urdu text. We use a grammatically motivated approach, which employs a sentiment-annotated lexicon. In this scheme, adjectives and adjectival phrases are combined with polarity shifters, and conjunctions to make sentiment expressions in the opinionated sentences. We label these sentiment expressions as, SentiUnits (Syed et al. 2010). For the identification and extraction of the SentiUnits from the given review we use shallow parsing based chunking. The classification algorithm works in combination with a sentiment-annotated lexicon of Urdu words. We evaluate the system using two corpuses of reviews from the domains of movies and electronic appliances.

The rest of the paper proceeds as follows: Section 2 gives a comprehensive overview of Urdu adjectival phrases in terms of morphology, classes and sentence structure. Section 3 describes the SentiUnits based sentiment classification model. Section 4 briefly presents the state of the art research in the field of sentiment analysis. Section 5 gives our experimentation and its results. Section 6 concludes our effort and suggests potential future endeavors.

2. URDU ADJECTIVAL PHRASES

In linguistics, for understanding the parts of speech (POS) of a language, we need to recognize their morphological structures and the processes through which these structures are made. Another significant aspect is to look at their different forms or classes. Therefore, we explore in this Section these two features of Urdu adjectival phrases. We first describe their morphological structures and then the classes.

2.1. Morphological Structure of Adjectives

The morphological structure of the Urdu adjectives is complex and exhibit frequent inflections and derivations with the agreement of the noun they qualify. Morphologically Urdu adjectives are categorized as unmarked and marked (Schmidt, 2000), (Omkar, 2008).

Unmarked adjectives: The unmarked adjectives do not show any inflection according to the nouns they qualify. In other words, they do not alter to show agreement with nouns through suffixes. Most of the Persian loan adjectives remain unmarked. For example the unmarked adjective; "دلچسپ" (*dilchasp*, interesting) remains unmarked with the nouns *(a)* masculine-singular, "ام " (*kaam*, task) as "دلچسپ ام" (*dilchasp kaam*, interesting task), *(b)* feminine-singular, " انا " (*khani*, story) as " انا دلچسپ" (*dilchasp khani*, interesting story) and (c) feminine-plural "انان " (*khanian*, stories) as "دلچسپ انان" (*dilchasp khaniian*, interesting stories).

Adjective marking: agreement in gender and number: The adjective marking is done through the suffixes for gender; masculine *(m)* and feminine *(f)* and for number; singular *(s)* and plural *(p)*. For example, the masculine adjective, "اچا" (*acha*, good) is inflected for gender as " اچی" (*achi*, good) and for number as " اچے" (*achay*, good). These suffixes are attached to agree with the noun or nouns, which the adjective qualifies. Therefore, there are three suffixes, i.e., singular-masculine *(a)*, singular-feminine *(ee)* and plural-masculine *(ay)*. Only one feminine suffix *(ee)* is used for singular and plural both.

Some examples of marked adjectives are given in Table 1, in this table, we have considered three nouns; *(a)* masculine-singular, " بچہ" (*bacha*, kid), *(b)* feminine-singular, "کار " (*car*, car) and masculine-plural "دن" (*din*, days). These noun cause inflection in the respective adjectives; "اچا" (*acha*, good), "لمبا" (*lamba*, long), and "برا" (*bura*, bad).

Table 1. Adjective marking with gender and number

Adjective *(m, s)*	**For gender *(f)***	**For number *(m, p)***
اچا بچہ (*acha bacha*, good kid)	اچی کار (*achee car*, good car)	اچے دن (*achay din*, good days)
لمبا بچہ (*lamba bacha*, tall kid)	لمبی کار (*lambee car*, long car)	لمبے دن (*lambay din*, long days)
برا بچہ (*bura bacha*, bad kid)	بری کار (buree *car*, bad car)	برے دن (*buray din*, bad days)

Agreement in case: Urdu nouns have three cases; oblique, nominative and vocative (Schmidt, 2000). All these cases cause an inflection in the adjectives. It means that the adjectives that qualify an oblique noun also become oblique. The masculine-singular suffixes *(a)* and *(an)* are replaced by, *(ay)* and *(ayn)*, respectively. The feminine adjectives remain the same as shown in Table 2. The adjectives "چھوٹا" (*chota*, little) and "ساتواں" (*satwan*, seventh) inflect with case.

Table 2. Adjective marking with gender and number

	Masculine	Feminine
Nominative	چھوٹا (*chota*) ساتواں (*satwan*)	چھوٹی (*chotee*) ساتویں (*satween*)
Oblique	چھوٹے (*chotay*) ساتویں (*satwayn*)	چھوٹی (*chotee*) ساتویں (*satween*)
Vocative	چھوٹے (*chotay*) ساتویں (*satwayn*)	چھوٹی (*chotee*) ساتویں (*satween*)

Adjectives with noun sequences: Some times adjectives appear in a sentence with more than one noun or multiple nouns making a sequence. In this case, the nouns may differ in gender and number. The adjective agrees with the noun, which is nearest to it. Examples are given in Table 3, in which, "بڑا" (*bara*, big) inflects for "پلنگ" (*palang*, bed) and "چھوٹی" (*choti*, younger) inflects for "خالہ" (*khala*, aunt).

Table 3. Adjective agrees with the nearest noun

Adjective	With the sequence of nouns
بڑا (*bara*, big)	بڑا پلنگ اور الماریاں (*bara palang aur almarian*, big bed and cupboards)
چھوٹی (*choti*,younger)	چھوٹی خالہ ، ماموں اوربچے (*choti khala, mamoon aur bachay*, younger aunt, uncle and kids)

Table 4. Adjective with partial and full reduplication

Partial Reduplication	Full Reduplication
ڈھیلا ڈھالا لباس (*dheela dhala libas*, loose dress)	بڑے بڑے کام (*baray baray kaam*, great tasks)
چھوٹی موٹی بات (*choti moti baat*, minute matter)	چھوٹی چھوٹی باتیں (*choti choti batain*, minute matters)

Reduplication of Adjectives: Urdu adjectives show reduplication either fully or partially. In full reduplication the whole word is repeated as it is, whereas, in partial reduplication some syllables of the word are reduplicated with different spellings. Examples of full and partial reduplication are given in Table 4.

2.2. Classes of Adjective

Urdu adjectives can be categorized as descriptive, predicative, attributive, possessive, demonstrative, and reflexive possessive, explained in following paragraphs:

Descriptive Adjectives: These are the most frequent and important type of adjectives. They describe attributes of the noun they qualify in terms of its size, dimensions, sound, color, shade, shape, quality, personal trait, or time, etc. Some examples of descriptive adjectives in Urdu are given in Table 5, where, "چھوٹا" (*chota*, little) and "لمبا" (*lamba*, long) describe the size of a noun, and "پیلا" (*peela*, yellow) and "سرخ" (*surkh*, red) express the color.

Table 5. Descriptive adjectives in Urdu

Category	Examples
Size	چھوٹا (*chota*, little) لمبا (*lamba*, long)
Color	پیلا (*peela*, yellow) سرخ (*surkh*, red)
Shape	مربع (*muraba*, square) تکونا (*tikona*, triangular)
Personal trait	اداس (*udaas*, sad) مجبور (*majboor*, helpless)
Qualities	مہربان (*mehrbaan*, kind) اچھا (*acha*, good)

Attributive Adjectives: If the descriptive adjectives directly precede a nominal head as modifiers then they are called attributive adjectives, because, they attributively modify or restrict the meaning of the noun. For example, the adjective "پیلا" (*peela*, yellow) modify the noun "غبارہ" (*ghubara*, balloon), to make it "پیلا غبارہ" (*peela ghubara*, yellow balloon). In this way the attributive adjective becomes part of the noun phrase. Some more examples are given in Table 6.

Table 6. Attributive adjectives modify the nouns

Nouns	Modified attributively
غبارہ (*ghubara*, balloon)	لال غبارہ (*laal ghubara*, red balloon)
چڑا (*chiria*, sparrow)	اداس چڑا (*udaas chiria*, sad sparrow)
بادشاہ (*badshah*, king)	نیک بادشاہ (*naik badshah*, kind king)

Predicative Adjectives: When the adjectives are used predicatively, they bring in new information about the noun instead of modifying it. These are not the component of the noun phrase, but are the complements of a copulative function, which links them to the noun, e.g., in Table 7, "غبارہ لال ہے" (*ghubara laal hay*, the balloon is red). In this case, the adjective "لال" (*laal*, red) identify the color of the noun "غبارہ"(*ghubara*, balloon). Only a specific feature of the noun is described both parts of speech, i.e., adjective and noun remain in their individual role. Some more examples are given in Table 7.

Table 7. Examples of predicative adjectives

Nouns	With predicative adjectives
غبارہ (*ghubara*, balloon)	غبارہ لال ہے(*ghubara laal hay*, balloon is red)
چڑا (*chiria*, sparrow)	چڑا اداس تھی (*chiria udaas thee*, sparrow was sad)
بادشاہ (*badshah*, king)	بادشاہ نیک ہے (*badshah naik hay*, king is kind)

Possessive Adjective: Possessive adjectives are used to indicate the possession. This possession relation is realized in two ways; whether, adjectives precede the head noun as modifiers in noun phrases like the attributive adjectives or they may be preceded by a suitable form of the genitive postposition "کا" (ka, of), "کے" (kee, of), and "کی" (kay, of). These genitive postpositions are lexically independent like "*of*" in English, but they agree in number and gender with the object noun. Consider the first example from Table 8, "ارتضیٰ کا پیلا غبارہ" (*Irtaza ka peela ghubara*, Itraza's yellow balloon). In this example the genitive postposition "کا" (ka, of) is used with a singular masculine noun, i.e., "پیلا غبارہ" (*peela ghubara*, yellow balloon). In the second example, "میرا" (*meri*, my) is a possessive adjective which is used for the first person and in this case is inflected for gender. Third example also contains the genitive postposition "کا" (ka, of) with a singular masculine noun.

Table 8. Examples of possessive adjectives

Examples
ارتضیٰ کا پیلا غبارہ (*Irtaza ka peela ghubara*, Itraza's yellow balloon)
میرا اداس چڑا (*meri udaas chiria*, my sad sparrow)
ایران کا مہربان بادشاہ (*Iran ka mehrbaan badshah*, kind king of Persia)

Demonstrative Adjective: The demonstrative pronouns act as the adjectives to indicate or demonstrate the specific inherent features of noun/ nouns of a particular type. As shown in Table 9, the Urdu demonstrative pronouns are different for *near* "ایسا" (*aisa*, like this), *far* "ویسا" (*waisa*, like that), *relative* "جیسا" (*jaisa*, such as) and *interrogative* "کیسا" (*kaisa*, how) demonstrations.

Table 9. Examples of demonstrative adjectives

Adjectives	Examples
ایسا (*aisa*, like this)	ایسا لباس (*aisa libas*, dress like this)
ویسا (*waisa*, like that)	ویسا لباس (*wasisa libas*, dress like that)
جیسا (*jaisa*, such as)	جیسا لباس (*jaisa libas*, such dress)
کیسا (*kaisa*, how)	کیسا لباس؟ (*kaisa libas*, what kind of dress)

Reflexive possessive adjective: The reflexive possessive adjectives are very frequently used in agreement with the noun they qualify, i.e., they inflect for gender, number and case. For example, "اپنا" (*apna*, own), "اس کا" (*uska*, someone else's) and "اس کا" (*iska*, someone else's) are used to indicate *one's own*, *someone else's far*, and *someone else's near*. The examples of the reflexive possessive adjective "اپنا" (*apna*, own) are given in Table 10, it is inflected for gender as "اپنی چابی" (*apni chabee*, one's own key) and for number as "اپنے لوگ" (*apnay loag*, one's own people).

Table 10. Examples of reflexive possessive adjectives

Nouns	With predicative adjectives
گ ر (*ghar*, house)	اپنا گ ر (apna *ghar*, one's own house)
چابۍ (*chabee*, kay)	اپنۍ چابۍ (*apni chabee*, one's own key)
لوگ (*loag*, people)	اپنے لوگ (*apnay loag*, one's own people)

Given this analysis we conclude that Urdu adjectival phrases are morphologically complex. In Section 2.1, we have discussed both marked and unmarked adjectives, which are borrowed from many languages, like Persian, Arabic, Hindi, Sanskrit, and English. This diversity results into flexibility and variety in the morphological and grammatical rules. For example, the adjectives which are Persian loan follow Persian grammar and usually remain unmarked, likewise, the Sanskrit based adjectives show inflections for gender and number, etc.

Similarly, some other linguistic phenomena are specific to Urdu language, e.g., frequent reduplication (partial as well as full) of adjectives, their inflection rules when used with a sequence of nouns. Almost all types of adjectives, descriptive, attributive, predicative, demonstrative, etc. show agreement in case, gender and number with the noun they qualify.

This linguistic behavior entails the need for the exact and well defined Urdu language specific grammar rules for the appropriate implementation and hence for getting better accuracy levels. We therefore, present a grammatically motivated approach for sentiment analysis of Urdu text.

3. SENTIMENT CLASSIFICATION USING ADJECTIVES AS HEAD WORDS

In our approach for sentiment classification based on adjectival phrases, we emphasize on the identification and extraction of the subjective expressions from the given text and coin a term "SentiUnits" for such expressions (Syed et al. 2010). The SentiUnits are the expressions made by single or multiple words, which are solely responsible for the whole sentimental orientation of a subjective sentence, i.e., a comment or an opinion. Our model is grammatically motivated and uses a sentiment-annotated lexicon for the identification of such expressions.

3.1. Examples of Adjective Based Subjective Expressions:

We take some examples of different types of adjectives discussed in Section 2, and discuss them in terms of SentiUnits, see Table 11. In the first sentence, a single predicative adjective "عمدہ" (*umdah*, good) represents the SentiUnit. The other possessive adjective "نئ" (*naee*, new) in the sentence is a part of the noun phrase. In sentence *b)* two predicative adjectives along with a conjunction in between, "ٹ ںڈا اور بدمزہ" (*thanda aur badmazah*, cold and tasteless) are recognized as the SentiUnit. In this case, the presence of conjunction "اور" (*aur*, and) augment the negative orientation of the sentence. The sentence *c)* expresses a strong positive orientation by a single attributive adjective in superlative structure " سب سے عمدہ" (*sab se umdah*, the finest). The positive adjective in the last sentence *d)*, "خوش رنگ" (*khoosh-rang*, of good color) joins with a polarity shifter "ن ہں" (*naheen*, not) to make a SentiUnit with negative polarity.

Table 11. Examples of sentences from Urdu

Example sentence
a) تمہ ار نئ تاب اور قلم عمدہ ہں- (*tmhari naee kitab aur qalam umdah hain*, Your new book and the pen are good.)
b) انا ٹ ںڈا اور بدمزہ - (*khana thanda aur badmazah hay*, The meal is cold and tasteless.)
c) سب سے عمدہ ناول - (*yeh sab se umdah novel hay*, This is the finest novel.)
d) لباس خوش رنگ ن ہں - (*libaas khoosh-rang naheen hay*, The dress is not of good color.)

a) Predicative adjective with sequence of nouns. Possessive adjective as part of noun phrase, *(positive)*
b) Multiple predicative adjectives with conjunction, *(negative)*
c) Single attributive adjective in superlative structure *(positive)*
d) Single predicative adjective with polarity shifter *(positive to negative)*

In most cases, the polarity shifters are the negation particles. In Urdu, both sentential and constituent negations exist (Omkar, 2008). The sentential negations are represented by the particles "مت" (*mat*, don't), "نا" (*na*, no) and "ن ہں" (*naheen*, not). Some frequently used constituent negations are "ن ہں" (*naheen*, not), "بنا" (*bina*, without), and "بغ ر" (*baghair*, without).

3.2. The Sentiment Analyzer Diagram:

Figure 1 shows the classification process of a given review. Each sentence in the review is analyzed one by one and for each sentence the polarity is calculated independently. Then these sentence polarities are combined to give review polarity.

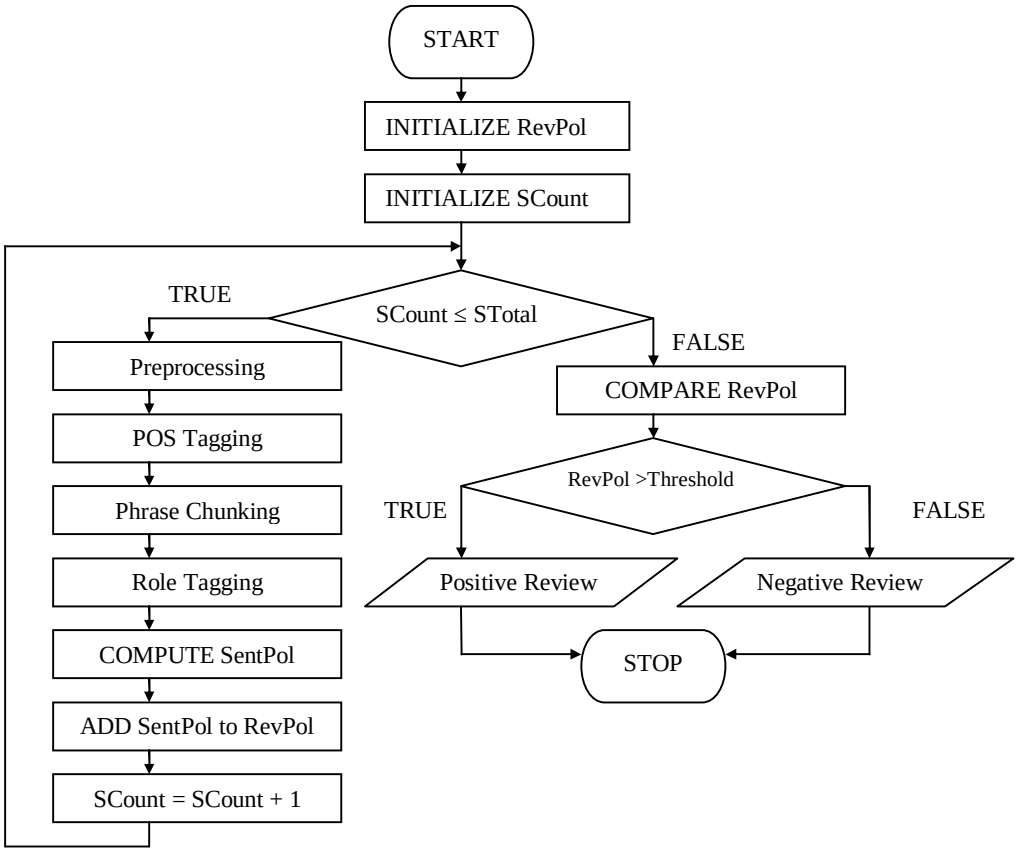

Figure 1. Sentiment classification of a review as positive or negative.

The algorithm initializes the review polarity *RevPol* and sentence count *SCount*. The sentiment analysis begins with the preprocessing of the given text. This step includes normalization, tokenization and finally the word segmentation. Urdu uses context sensitive script and therefore the tokenization and word boundary identification are handled separately (Syed et al. 2010). The preprocessed words are then assigned parts of speech tags, e.g., nouns, verbs, adjectives, conjunctions, and negations etc. These tagged words are converted into phrases by phrase chunking. Consequently, we get noun phrases, verb phrases, and adjective phrases etc.

The algorithm takes the adjectives as the head words and looks for the polarity shifters, and conjunctions to assemble them as SentiUnits in the opinionated sentences. As a result role tagging is done and the complete SentiUnits with attached polarity shifters and conjunctions are identified. The algorithm compares and then computes the polarity values for each SentiUnit, whereas, the sentence polarity *SentPol* is computed by the total polarities of its constituent SentiUnits. To calculate the review polarity *RevPol*, all the known sentence polarities *SentPol* are added and compared to the threshold value. If the result is greater than threshold then, the review is positive and vise versa.

3.3. Example:

As shown in table 12, we take an example of Urdu sentence and apply shallow parsing based chunking to extract the SentiUnits. As the Figure 1, depicts, this extraction is achieved in three steps, *a)* assign parts of speech tags to all preprocessed words, *b)* identify phrases through phrase chunking, and *c)* extract SentiUnits through role tagging. The given sentence contains a complex SentiUnit made by a positive descriptive/predicative adjective and a polarity shifter.

"ارتضا اورفاطمہ اسکول شادہ نہ ہیں-"

(*Irtaza aur fatima ka skool kushadah naheen,* Irtaza and fatima's school is not spacious)

Table 12. Shallow parsing based chunking of the given sentence.

Parse of the complete sentence	[N CJC N PM N] [ADJ NEG] à NP SU	ارتضہ اورفاطمہ کا سکول شادہ نہیں
Parse of the Noun phrase with possession marker (PM) and conjunction (CJC)	N CJC N PM N à NP	ارتضہ اورفاطمہ کا سکول
Parse of the SentiUnit with negation (NEG)	ADJ NEG à SU (SentiUnit)	شادہ نہیں

The SentiUnit "شادہ نہیں" (*kushadah naheen*, not spacious) contains an adjective head and a negation word. The noun phrase in the opinion is even more complex, i.e., "ارتضہ اورفاطمہ کا سکول" (*Irtaza aur fatima ka school*, Irtaza and fatima's school).

We give the adjectives based sentiment classification analyzer in Figure 1, the system identify and extract the adjective based subjective expressions called SentiUnits and on the basis of their polarity it classy each sentence one by one. The final classification of the review is an accumulation of all computed scores. The working of the analyzer along with an example clearly explains our approach. Next, we evaluate this approach using a lexicon and two corpora of the reviews.

4. SENTIMENT ANALYSIS RESEARCH

The part of speech based features of the given text, particularly of adjectives, can help a lot in sentiment analysis. That is why one of the earliest works in this domain (Hatzivassiloglou & McKeown, 1997) uses adjectives as subjectivity indicators. They employ a log-linear regression model for identification and validation of the positive or negative semantic orientation of the conjoined adjectives. A clustering algorithm divides the adjectives into groups with respect to orientations, and labels them as positive or negative. Before that (Hatzivassiloglou & McKeown, 1993), present an approach for automatic recognition of adjectival scales this approach group or cluster the adjectives carrying same semantics, but this was not with the perspective of sentiment analysis. (Bruce & Wiebe, 2000) recognize subjectivity within the text by manual tagging. They take a case study of sentence level categorization and categorize clauses from the "Wall Street Journal" as objective or subjective. Each clause is given a final classification on the basis of an agreed decision by four judges.

(Hatzivassiloglou & Wiebe, 2000) analyze two main features of adjectives for subjectivity prediction, i.e., gradability and semantic orientation. They extract reliability of gradability values using an automatic method for extracting. (Turney, 2002), suggest that the proverbs are also carriers of sentiments in a sentence and should be considered in combination with adjectives. In their work, the sentences are divided into pre-structured grammatical patterns, which include adjectives and adverbs as the core word. (Riloff et al., 2003) emphasize on the identification of the subjective nouns, which are modified by the use of adjectives. They compute the orientation of the phrases in the sentence that contained them. (Riloff & Wiebe, 2003), use unsupervised learning method for automatic extraction and learning of the patterns for subjective expressions in the given text.

(Whitelaw et al., 2005) propose the use of appraisal theory for sentiment analysis. They work on appraisal expressions extraction. These appraisal expressions are the sentiment oriented phrases which contain adjectives as head words. (Bloom & Argamon, 2010) extended this model and propose an approach for automatic learning of these appraisal expressions. Research contributions related to adjective based sentiment analysis are shown in Table 13.

Table 13. Research contributions related to adjective based sentiment analysis

Feature	Example Contributions
Adjectives	Hatzivassiloglou, McKeown (1993) Hatzivassiloglou, McKeown (1997) Bruce and Wiebe (2000) Hatzivassiloglou, Wiebe (2000)
Adjectives and proverbs	Turney (2002)
Subjective Nouns	Riloff et al. 2003
Expressions	Riloff and Wiebe (2003) Whitelaw et al. (2005) Bloom and Argamon (2010)

5. EXPERIMENTATION

The performance evaluation of the NLP based classifiers is done through experimentation. Experiments are performed with the help of lexicons applied on the test-beds (corpuses). In sentiment classification, the domain specific lexicons and corpuses exhibit variations in results. Thus, the task of evaluation of the sentiment classifier itself becomes a major concern. Our lexicon and corpuses are briefly described below:

5.1. Sentiment- Annotated Lexicon and Corpus:

Lexicon construction with appropriate coverage is a challenging and time consuming task. There are a number of efforts (Hatzivassiloglou & Wiebe (2000), Turney (2002), (Riloff et al., 2003) and (Higashinaka et al., 2007) which have tried to develop algorithms and techniques for automatic lexicon construction using unsupervised learning methods. Some other contributions have tried to use or extend the existing lexicons, e.g. the extension of WordNet is SentiWordNet. (Hu & Liu, 2005), (Andreevskaia & Bergler, 2005), and (Annett & Kondrak (2008) use WordNet or its extension for sentiment analysis.

Most of these efforts are for English language and use already prepared linguistic recourses like corpuses for automatic extraction of required lexicons. Therefore, for English language this facet of research is no more an unresolved issue. But for a recourse poor language (Muscand & Ghosh, 2010) like Urdu this task poses many challenges. Therefore, for experimentation we use our manually constructed lexicon of Urdu words. All the entries are marked with sentiment polarity values.

We have collected two corpuses as the test-beds from the domains of movies and electronic appliances. The movie reviews based corpus *C1* contains 450 reviews, among which 226 are positive and 224 are negative. For obtaining diversity in opinions we present 20 different movies to the reviewers, from three categories, i.e., action, comedy, and horror. The corpus containing reviews of electronic appliances *C2* includes 328 reviews, among which 177 are positive and 151 are negative. The electronic appliances presented for review are refrigerators, air-conditioners and televisions, taken from three different brands.

5.2. Results

For result generation we have applied accuracy *A* as the classification performance metric. Two experiments are performed on corpora *C1* and *C2*. We also analyze the behaviors of positive and negative reviews separately. Table 14, shows the results, with accuracy of 66-74% for *C1* and 77-79% for *C2*.

Table 14. Experimental results in terms of accuracy

Orientation	Corpora	A
Negative	*C1*	66%
	C2	77%
Positive	*C1*	74%
	C2	79%

Comparison between the results from both corpora is given in Table 15. The total accuracy of the model is 74%.

Table 15. Comparison of accuracy from both corpora *C1* and *C2*

	Accuracy				
C1 *Neg*	66%	8%	70%	8%	74%
Pos	74%				
C2 *Neg*	77%	2%	78%		
Pos	79%				

From the above results it is clear that the accuracy of a sentiment classifier is domain dependent. The movie reviews show lower accuracy (accuracy difference = 8%) than the electronic appliances. This variation is due to the difference in the complexity levels of both domains. This is noted that the appliances reviews are simpler and to the point as compared to those of movies. In movie reviews the target of opinion or comment is not always the movie but it can be the characters, story, or the direction etc. Another observation from the results is that the rate of misclassification for negative reviews is more as compared to positive. The main reason for this misclassification is the use of polarity shifters which sometimes can cause erroneous results.

6. CONCLUSION

This research work gives a comprehensive analysis of morphology, grammar and structure of Urdu adjectives and adjectival phrases, and concludes that these are morphologically complex and follow flexible grammar rules due to the extendable vocabulary of the language. We take more perceptible features of adjectival phrases and then write the rules for accurate extraction of these phrases. These features include, sentence structures containing these phrases, their position with respect to nouns, use of postpositions, morphological changes, reduplication, etc. As the next step we

extract the SentiUnits which, are made by adjectives/ adjectival phrases, conjunctions and polarity shifters. This extraction is achieved by implementing shallow parsing based chunking. Despite of inherent complexities of the language we achieve excellent results of this effort (74%).

Our future focus is on the same SentiUnits based approach. We plan to attach these expressions with the candidate targets. These targets are the noun phrases. For this purpose we need to extend the classification model for noun phrase identification as well as the lexicon.

7. REFERENCES

- Afraz Z. Syed, Aslam Muhammad, Ana María Martínez Enríquez: Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. MICAI (1) 2010: 32-43
- Andreevskaia, A., Bergler, S.: Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In: EACL 2006, Trent, Italy (2005)
- Annet, M., Kondark, G.: A comparison of sentiment analysis techniques: Polarizing movie blogs. In: Bergler, S. (ed.) Canadian AI 2008. LNCS (LNAI), vol. 5032, pp. 25–35. Springer, Heidelberg (2008)
- Bloom, K., Argamon, S.: Unsupervised Extraction of Appraisal Expressions. In: Farzindar, A., Kešelj, V. (eds.) Canadian AI 2010. LNCS (LNAI), vol. 6085, pp. 290–294. Springer, Heidelberg (2010).
- Durrani, N., Hussain, S.: Urdu Word Segmentation. In: 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Los Angeles, US (2010).
- E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.
- E. Riloff, J. Wiebe, and T. Wilson, “Learning subjective nouns using extraction pattern bootstrapping,” in Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 25–32, 2003.
- Hu, M., Lui, B.: Mining and summarizing customer reviews. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (2005)
- Koul N. Omkar: Modern Hindi Grammar, Dunwoody Press (2008).
- Muaz A., Ali A. and Hussain S., Analysis and Development of Urdu POS Tagged Corpora, In the Proceedings of the 7th Workshop on Asian Language Resources, IJCNLP’09, Suntec City, Singapore, 2009.
- Mucand, S., Ghosh, D.:Using Cross-Lingual Projections to Generate semantic Role Labeled Corpus for Urdu- A Resource Poor Language. In: 23rd International Conference on Computational Linguistics (Coling 2010
- P. Turney. 2002. Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In EMNLP.
- R. Higashinaka, M. Walker, and R. Prasad, “Learning to generate naturalistic utterances using reviews in spoken dialogue systems,” ACM Transactions onSpeech and Language Processing (TSLP), 2007.
- Rebecca Bruce and Janyce Wiebe. 2000. Recognizing subjectivity: A case study of manual tagging. Natural Language Engineering, 6(2).
- Riaz, K.: Challenges in Urdu Stemming. Future Directions in Information Access, Glasgow (August 2007).
- Schmidt, R.: Urdu: An Essential Grammar. Routlege Publishing, New York (2000).
- Vasileios Hatzivassiloglou, Kathleen McKeown: Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. ACL 1993: 172-182.
- Vasileios Hatzivassiloglou, Kathleen McKeown: Predicting the Semantic Orientation of Adjectives. ACL 1997: 174-181.
- Vasileios Hatzivassiloglou and Janyce Wiebe. "Effects of Adjective Orientation and Gradability on Sentence Subjectivity". In Proceedings of the 18th International Conference on Computational Linguistics (COLING-00), Saarbrücken, Germany, August 2000.
- Whitelaw, C., Garg, N., Argamon, S.: Using appraisal taxonomies for sentiment analysis. In: SIGIR (2005).

21/03/2011