Mining the Urdu Language-Based Web Content for Opinion Extraction

Afraz Z. Syed[1], A.M. Martinez-Enriquez[2], Akhzar Nazir[3], Muhammad Aslam[3(✉)], and Rida Hijab Basit[3]

[1] Information Technology Program (ITP), Lambton College of Applied Science and Technology, Sarnia, Canada
Afraz.Syed@lambtoncollege.ca
[2] Department of CS, CINVESTAV-IPN, D.F. Mexico, Mexico
ammartin@cinvestav.mx
[3] Department of CS and E, University of Engineering and Technology, Lahore, Pakistan
akhzarn@yahoo.com, maslam@uet.edu.pk, ridahijab@gmail.com

Abstract. People prefer to share and express opinions in their own language. Internet is a biggest repository for sharing opinions. Opinion mining uses Natural Language Processing (NLP), text analysis and computational linguistics to identify and extract subjective information in data. Opinion mining for Urdu language is not a well explored area. Therefore, an approach has been proposed which identifies and extracts adji-units and decisions from the given text using lexicon-based approach focusing on Urdu language. Adji-units are the expressions which contain subjective text in a sentence. Our proposed approach uses two-step lexicon to extract opinions from text chunks. Moreover, for Urdu language no such lexicons exist. The main aim is to develop a diverse two-step lexicon and highlight the linguistic as well as technical aspects of this multi-dimensional research problem. The performance of the proposed system is evaluated on multiple texts and the achieved results are quite satisfactory.

Keywords: NLP · Opinion mining · Sentiment analysis · Urdu lexicon · Adji-units

1 Introduction

World Wide Web has emerged as the largest repository of the user generated texts consisting of opinions. Suggestions from different people exist on the Internet and Internet users, nowadays, use forums, news blogs, discussion groups or review sites for opinions and suggestions even while taking a smallest decision e.g. buying a routine device [1]. The opinions can be defined as the subjective expressions that describe people's feelings [2], sentiments, or appraisals towards objects, procedures, events and their characteristics.

Sites, forums or discussion groups gathering opinions consist of bulks of data making difficult for a person to search for relevant opinions manually. Moreover,

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2017, LNCS 10267, pp. 244–253, 2017.
DOI: 10.1007/978-3-319-59226-8_24

survey companies may also need such data to carry out a research about any product, person or political party. Hiring individuals for this job would be costly and time consuming. Therefore, a system is needed which automatically mines through such data to get relevant opinions or suggestions about any specific thing.

Different approaches exist for opinion mining like supervised, unsupervised, lexicon based approaches. These techniques have been used for mining opinions of different languages like English, Persian, Hindi, Turkish but none has been used for Urdu language. Our proposed system focuses on Urdu language as it is a major language spoken and understood around the globe with 80 million speakers in the subcontinent [3]. It poses certain challenges due to its complex morphology and orthography. These challenges have to be overcome during Urdu language processing.

Different forums and social media sites are localizing their content by allowing users to comment and chat in their native language. Such forums are tremendously increasing for Urdu language as well where people can add their suggestions in Urdu. Due to this, we have proposed an approach for Urdu language which analyzes the data and highlights positive and negative opinions. In this work, Lexicon-based approach has been implemented.

Section 2 reviews related work in opinion mining, Sect. 3 briefly describes the implementation of two-step lexicon based opinion mining model for Urdu (LOMMU), Sect. 4 discusses the evaluation results of LOMMU, whereas, Sect. 5 concludes the paper along-with some future directions.

2 Literature Survey

Many different approaches for opinion mining have been proposed by different researchers. Learning methods that are supervised, unsupervised, and semi-supervised in nature have been used by some of them. Unsupervised learning methods have been increasingly successful in recent NLP research mainly because it takes unlabeled data as input. Moreover, unsupervised learning results in better understanding of modeling methods, optimization of algorithms and conversion of domain knowledge into structured models. Sentiment analysis also uses lexicon based approach with un-supervised learning method. Three different approaches are used to construct sentimental lexicon - manual, dictionary-based or corpus-based approach.

Naïve Bayes algorithm is the most widely used supervised classification model [4]. It estimates the probabilities of opinions (as positive or negative) using the joint probabilities of a set of words in a given category. Support Vector Machine (SVM) is a non-probabilistic binary classification method proposed by Vladimir Vapnik. It looks for a hyper plane with the maximum margin between positive and negative examples of the training opinions. In addition to the above, K-Nearest Neighbor (KNN) classification (KNN) is based on the assumption that the classification of an instance is most similar to classification of other instances that are nearby in the vector space. In comparison to the other classification methods such as Naïve Bayes, KNN does not rely on prior probabilities and is computationally efficient [5]. Naïve Bayes, SVM, and KNN classifiers discussed above have been used for English language opinion mining. All these are termed as supervised learning methods. Another technique has been proposed which

performs classification based on some fixed syntactic patterns that are likely to be used to express opinions [6]. Lexicon-based approaches have also been used by English language. Comprehensive lexicons have been constructed for English language like SentiWordNet 3.0 which is publically available and is used by different researchers for opinion extraction [7].

Cross-domain sentiment analysis has been done by many researchers [8] experimented with German emails. German emails are converted to English for calculating sentiment orientations, after which they are again converted to German. Precision of this system is satisfactory but recall is recorded to be quite poor. In [9], a slightly different problem has been attempted by using a maximum entropy-based EM algorithm. It jointly learns two monolingual sentiment classifiers by treating the sentiment labels in the unlabeled parallel text as unobserved latent variables.

Urdu is a morphologically complex language having a different writing style due to which using cross-domain sentiment analysis technique for Urdu opinion mining would be quite difficult. Moreover, Urdu data available online is unlabeled and less data is available for analysis. Therefore, less work related to lexicon implementation has been done using corpus. One of the most comprehensive Urdu language lexica is available at http://www.cle.org.pk [3]. This data is XML based, as per the annotation schema, containing about 20 etymological, phonetic, morphological, syntactic, semantic and other parameters of information about a word. Another lexicon proposed by [10] has been constructed from news Urdu corpus having 1.5 million words. It has been tokenized on space and punctuation marks, keeping the diacritics. Extracted lexicon contains 9,126 total words and 4,816 unique words. These lexicons do not contain enough data for decision making. Moreover, accuracy of these lexicons is not as good as described by the researchers.

Urdu lexicon development involves decisions regarding parts-of-speech (POS) tags and their respective features, lemmas, transcription, and lexicon format. POS tagger used for Urdu lexicon development tags sentence on the basis of noun, verb, adjective, adverb, numeral, postpositions, conjunctions, pronouns, auxiliaries, case markers, harf, etc. Most of the on-going works have used XML based lexicon formats [11, 12]. Construction of such lexicons is time consuming as each scenario has a detailed information attached to it.

An alternate solution to XML based would be a Java Script Object Notation (JSON) based two-step lexicon approach as it is easy to implement and is less time consuming. It consists of different keys and each key has corresponding values associated with it. Proposed lexicon-based opinion mining model for Urdu (LOMMU) using JSON format is described in the next section.

3 Lexicon-Based Opinion Mining Model for Urdu Language (LOMMU)

Developing a lexicon for opinion mining is quite critical. LOMMU can be implemented for any operating system (OS) but our work has been tested with Macintosh OS. It uses an algorithmic approach to develop a two step lexicon. JSON format based lexicon structure is shown in Fig. 1.

```
{
  "Orthography": "مردوں",
  "ENTRY": [
    {
      "NOM": {
        "Case": "oblique",
        "Number": "plural",
        "Gender": "masculine"
      },
      "LEMMA": "مرد",
      "PHONETIC": "m @ r - d_d o~"
    },
    {
      "NOM": {
        "Case": "oblique",
        "Number": "plural",
        "Gender": "invariant"
      },
      "LEMMA": "مردہ",
      "PHONETIC": "m U r - d_d o~"
    }
  ]
}
```


Fig. 1. JSON Format-based Lexicon Structure

JSON format given above gives detailed information about the word ""مردوں"" {Mardon, Men}. It contains gender information, number, phonetics, case, and lemma of the candidate word.

Raw corpus has been annotated using a POS tagger and then adji-units are extracted from the given text. All the decisions which are made during opinion mining use adji-units. Negations have also been handled in our system by using them as polarity shifters. LOMMU uses a two-step lexicon consisting of positive and negative lexemes. Extracted adji-units from the text under consideration are compared with the lexemes and in case of negations attached with adji-units, the polarity of the sentence shifts. System overview has been given in Fig. 2.

Here, we define our problem of Urdu opinion mining. Let "O" be the Urdu text consisting of sentences which can be factual or opinionated. So we can say that "O" is a union of factual and opinionated sentences:

$$O = \{\text{set of factual sentences}\} \cup \{\text{set of opinionated sentences}\}$$

LOMMU differentiates opinionated sentences from factual ones because of the significance of opinionated sentences in opinion mining. The main tasks of LOMMU can be described as follows:

- **Convert Gathered Data into UTF-16 Format for Processing:** Gathered data is in different forms and hence, cannot be processed as it is. Therefore, it is converted to UTF-16 format for further processing.

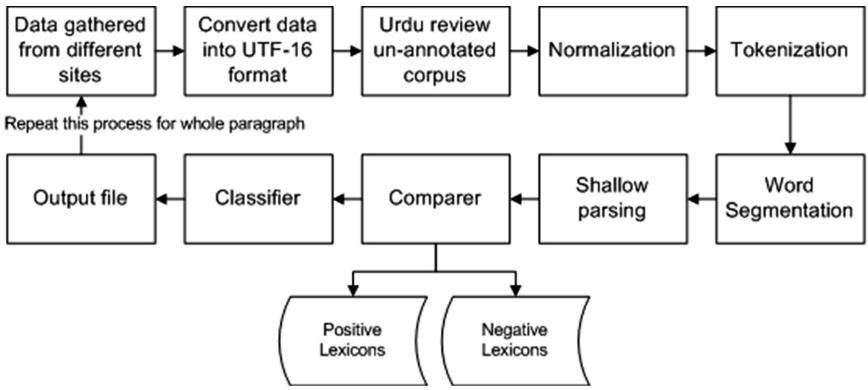

Fig. 2. Two Step Lexicon based Opinion Mining Model for Urdu Language

- **Normalization of Data:** The converted data is then normalized by removing dots, punctuation marks, commas, dashes or any irrelevant symbols. This step can be referred to as a preprocessing step.
- **Tokenization:** Extracting each word from a sentence is known as the process of tokenization. Tokens are just separated by spaces and may not be complete meaningful words. Some examples of tokens for the following sentence are given below:

ایپل کے کمپیوٹرز بہت خوبصورت ہوتے ہیں {Apple computers are very beautiful}

<Token>	ایپل {Apple}	<Token>
<Token>	کمپیوٹرز {computers}	<Token>
<Token>	بہت {very}	<Token>
<Token>	خوبصورت {beautiful}	<Token>
<Token>	ہوتے {are}	<Token>

- **Segmentation:** It is a process of extracting meaningful words. Some tokens are not meaningful words as said in the previous step; therefore, segmentation is needed to get a complete meaningful segment of a word. Examples of segments for the same sentences given above are shown below:

<Word>	ایپل {Apple}	<Word>
<Word>	کمپیوٹرز {computers}	<Word>
<Word>	بہت {very}	<Word>
<Word>	خوبصورت {beautiful}	<Word>

- **Shallow parsing:** Adji-units are extracted for opinion mining using shallow parsing after annotating the corpus. Any POS tagger e.g. CRULP POS tagger can be used for annotating the corpus. Phrase level negations are also handled as part of shallow parsing. Examples of shallow parsing are given below with reference to the sentence given above.

<NP>	ایپل کے {Apple's}	<NP>
<Noun>	کمپیوٹرز {computers}	<Noun>
<Verb>	ہوتے {are}	<Verb>
<Adji-unit>	بہت خوبصورت {very beautiful}	<Adji-unit>

- **Adji-units Analysis:** Adji-units are then compared with the positive and negative lexemes in the lexicon. Due to this, it is known as two-step lexicon based opinion mining model for Urdu language. The presence of the word (بہت {very}) in a sentence enhances the intensity of that sentence (either positively or negatively). Overall polarity of the sentence is then calculated. Adji-units which do not match with either positive or negative entries in the lexicon have been entered manually for efficient processing. lexicons.

4 Evaluation of LOMMU

LOMMU has been evaluated by using sample text files consisting of sentences and 10,000 tagged words downloaded from http://www.cle.org.pk. First of all, tag-set has been selected to extract adji-units. Secondly, extracted adjectives have been compared with positive and negative lexemes in the lexicon. Finally, results have been discussed along-with the system accuracy. Figures 3 and 4 show complete working of LOMMU.

Developed LOMMU reads a text file for which polarity has to be calculated. It extracts the list of adji-units from the tagged words by retaining the words with a tag <JJ> . Negations and other factors that may increase or decrease polarities are stored at backend to be used while calculating final results. Extracted adji-units are then compared with the entries in the lexicon and sentence-by-sentence analysis is conducted for making the overall decision.

Fig. 3. Read Urdu Tagged File, Tag-set Selection and Adji-units Extraction

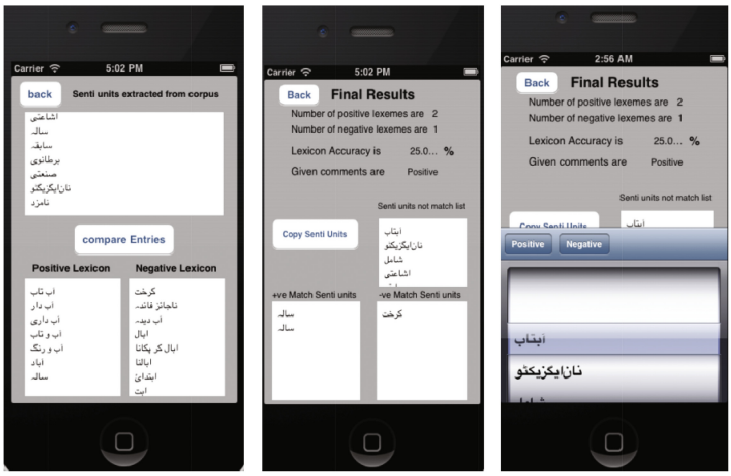

Fig. 4. Lexicon having Negative and Positive Lexemes; and Extracted Adji-units

Figure 3 shows adji-units extracted from the given file and results given by LOMMU.

LOMMU has been tested with test data and results obtained are satisfactory. Some sample texts have been discussed here along-with the results given by our system when these texts are passed through it.

Sample 1: ایپل کمپیوٹرز {Apple Computers}

پچھلے مہینے میں نے ایک لیپ ٹاپ خریدا ہے. یہ ایک شاندار رفتار بہت حیرت انگیز ہے. آپریٹنگ چیز ہے. اس کی پروسیسنگ سسٹم بہترین ہے. اگرچہ بیٹری دیرپا نہیں ہے، جو میرے لئے قابل قبول ہے. یہ لیپ ٹاپ دیکھنے میں بھی بہت خوبصورت ہے. یہ ایپل کمپنی کا کمپیوٹر ہے. اس کمپیوٹر میں وائرس آنے کا خطرہ نہیں ہے. یہ کمپیوٹر بہت مہنگا ہے. ہر شخص اس کمپیوٹر کو نہیں خرید سکتا. میں ایک سوفٹ ویئر انجینئر ہوں اور میرے لئے اس کمپوٹر کو خریدنا بہت ضروری ہے. ایپل کے لیپ ٹاپ کو میک بک بولا جاتا ہے. اب ایپل کمپنی نے مختلف رنگوں میں میک بک متعارف کروائی ہیں. اس کی وجہ سے ایپل کمپیوٹر کی مارکیٹ بڑی تیزی سے آگے بڑھ رہی ہیں. اس کمپنی کی جتنی بھی پروڈکٹس ہیں وہ اچھی اور پائیدار ہیں. جن آئی فون سر فہرست ہیں.

{I bought a laptop last month. It is a wonderful thing. Its processing speech is quite astonishing. Operating system is wonderful. Although battery is not long-lasting, but it is okay for me. This laptop looks very beautiful. It belongs to Apple company. This computer is not endangered to any virus. This computer is very expensive. Everyone can buy this computer. I am a software engineer and it was very important for me to buy this computer. Apple's laptop is called as Mac book. Now, Apple company has introduced Mac books in various colors. Due to this, Apple computer's market in increasing rapidly. All the products of this company are good and long-lasting. Iphone is one of them.}

Sample 1 contains reviews about laptop taken from a discussion forum. More than 400 words have been minimized to around 160 words making a complete paragraph. This data has then been tagged using an existing POS tagger. LOMMU has read data word by word and extracted adji-units as shown by underlined words in the text. Negations attached with any of the extracted adji-unit have been stored at backend for final decision-making.

Table 1. Sample 1 results

Sample 1 - Final results
System extracts 4 out of 8 positive lexemes and 2 out of 2 negative lexemes. So, total matched lexemes are 6 out of 10. Therefore, lexicon accuracy is 60%. After using negations as polarity shifters, the overall opinion about given data is positive.

When we pass this text from our system it matches 4 positive lexemes and 2 negative lexemes but skips others. Positive lexemes have been shown by simple underlined words whereas negative lexemes have been shown with dotted underlined words. Skipped lexemes can be manually added into the existing lexicon list for future use. Overall results for sample 1 have been discussed in Table 1 below. For the given text in sample, the accuracy of LOMMU lexicon is 60%.

Sample 2: میرا دوست{Mera Dost}

میرے بہت سے دوست ہیں۔ لیکن محسن میرا سب سے اچھا دوست ہے۔ محسن ایک ڈیزائنر ہے اور ایک بہت بڑی کمپنی میں بہت اعلیٰ درجے پر فائز ہے۔ محسن ایک اعلیٰ تعلیم یافتہ نوجوان ہے۔ محسن ہمیشہ سچ بولتا ہے اور اپنا کام محنت اور لگن سے کرتا ہے۔محسن کے دفتر میں ہر شخص اس کی شخصیت سے متاثر ہے۔ محسن اپنے زمانہ طالب علمی سے ہی بہت محنتی ہے۔ محسن ہمیشہ میرے مشکل وقت کا ساتھی ثابت ہوا ہے۔ وہ کھانے میں اکثر بے احتیاطی کرتا ہے جسکی وجہ سے وہ اکثر بیمار رہتا ہے اس کی اس عادت کی وجہ سے اس کے دوست اور گھر والے بہت تنگ ہیں۔ محسن نے کبھی اپنی صحت کا خیال نہیں رکھا۔ وہ اگر محنت کے ساتھ ساتھ اپنی صحت کا بھی خیال رکھے تو وہ مزید خوشگوار زندگی گزارسکتاہے۔ کیونکہ صحت مند جسم ہی صحت مند دماغ ہوتا ہے۔

{I have many friends. But Mohsin is my best friend. Mohsin is a designer and works in a very big company at a very good post. Mohsin is a well-educated young man. Mohsin always speaks truth and does his work with dedication and passion. Everyone in Mohsin's office is impressed by his personality. Mohsin is very hard-working from his student life. Mohsin has always proven to be my partner in my difficult times. He often shows carelessness in eating due to which he always remains ill. His friends and family members are fed up of this habit of his. Mohsin has never taken care of his health. If his takes care of his health along-with the hardwork he does, he can live a more happy life. Because, healthy body is a healthy mind.}

Sample 2 discusses reviews about an employee of a software firm. Here, 350 words have been reduced to 250 words making a complete paragraph. This sample data has also been converted to tagged data using POS tagger. Extracted adji-units have been

shown by the underlined words in the text above. Simple underlined words are positive lexemes whereas dotted underlined words are negative lexemes.

Sample 2, when passed through our system, matches 8 positive lexemes and 2 negative lexemes. In this case, the overall accuracy of LOMMU lexicon is recorded as 55.55% which has been given in Table 2 below.

Table 2. Sample 2 results

Sample 2 - Final results
LOMMU extracts 8 out of 14 positive lexemes and 2 out of 4 negative lexemes in case of sample 2. Therefore, lexicon accuracy is 55.55% as 10 out of 18 lexemes have been matched. After using negations as polarity shifters, overall opinion about given data is positive.

Entire experimentation has been conducted using different corpuses having around 100,000 words. This experiment has given a decreased accuracy of 50–52%. The main reason for this accuracy decline is that our lexicon is not mature enough and contains only 15,000 words (adji-units). Adji-units can be added manually for efficient processing. Increasing the number of adji-units in the lexicon would definitely increase the LOMMU accuracy.

LOMMU presents a sentiment-annotated lexicon for mining opinionated positive and negative expressions of any given Urdu text. It is an integral basis of Urdu text based sentiment analysis. LOMMU gives an accuracy of about 50–52% with just 15,000 adji-units in the lexicon. Increasing this further would definitely increase the LOMMU lexicon accuracy.

Moreover, all the existing sentiment analysis systems are for Windows platform. Our system, on the other hand, provides a platform for Macintosh users.

5 Conclusion and Future Work

Two-step lexicon based opinion mining model has been proposed for Urdu language which uses a JSON based approach for constructing the lexicon. For each word in the lexicon, detailed information has been given. It has been tested with different corpuses having about 100,000 words. The system gives an accuracy of about 50-52% as our lexicon consists of only 15000 words.

Future work associated with it would be the enhancement of developed lexicon by adding more words so that high system accuracy can be achieved.

References

1. Syed, A.Z., Aslam, M., Martinez-Enriquez, A.M.: Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu Text. Artif. Intell. Rev. **41**(4), 535–561 (2014)
2. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of 12th International Conference on World Wide Web, pp. 519–528 (2003)
3. Hussain, S.: Resources for Urdu language processing. In: Proceedings of 6th Workshop on Asian Language Resources IJCNLP, pp. 1–10 (2008)
4. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. J. 1138–1152 (2011)
5. Han, E.H.S, Karypis, G., Kumar, V.: Text categorization using weight adjusted k-nearest neighbor classification. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 53–65 (2001)
6. Turney, P.D.: Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002)
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
8. Kim, S.M., Hovy, E.: Identifying and Analyzing Judgment Opinions, In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 200–207 (2006)
9. Lu, B., Tan, C., Cardie, C., Tsou, B.K.: Joint bilingual sentiment classification with unlabeled parallel corpora. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics – Human Language Technologies, vol, 1, pp. 320–330 (2011)
10. Humayoun, M., Hammarström, H., Ranta, A.: Urdu morphology, orthography and lexicon extraction. In: Proceedings of 2nd Workshop on Computational Approaches to Arabic Script-based Languages (2007)
11. Ijaz, M., Hussain, S.: Corpus based Urdu lexicon development. In: Proceedings of Conference on Language and Technology (CLT), pp. 1–10 (2007)
12. Syed, A.Z., Muhammad, A.: Lexicon based sentiment analysis of Urdu text using senti-units, In: Proceedings of 10th Mexican International Conference on Advances in Artificial Intelligence, pp. 32–43 (2010)