A Sentiment Lexicon for Urdu

Sahar Rauf, Kinza Rahim, Maryam Khalid, Ehsan ulHaq, Kashif Javed
Center for Language Engineering, Al-Khwarizmi Institute of Computer Sciences University of Engineering and Technology Lahore, Pakistan
firstname.lastname@kics.edu.pk

Abstract

Sentiment analysis is a data mining technique, which measures the inclination of people's opinions. Recent studies have shown that the sentiment lexicon can be developed using automatic and manual tagging techniques. The seminal works on Urdu lexicon done so far do not actually denote a broad Lickert scale for data tagging and also do not cover all the open word classes. The current study aims to develop a sentiment lexicon and test its validity using manual and automatic methods. The dictionary-based method is used to design this lexicon using three authentic Urdu dictionaries. The data was tagged on a five point lickert scale i.e. -2 to +2 using the formulated guidelines. The lexicon is composed of four-word classes namely nouns, verbs, adjectives and adverbs. Once the lexicon was developed using manual tagging techniques it was tested both manually and automatically. The manual testing yielded an inter annotator agreement of 75% while the automatic testing included the comparing of the sentiments of the developed lexicon with the UCI Corpus. The result yielded a percentage accuracy of 84%. The lexicon was validated for its accuracy by both the results.

1. Introduction

Sentiment analysis (SA) is one of the fastest growing fields of Natural Language Processing (NLP) and text mining under the umbrella of Artificial Intelligence (AI) [1]. SA measures the inclination of people's opinions through NLP, computational linguistics and text analysis. Sentiment analysis has emerged from human behavior of decision making by consulting friends or family about their opinions in daily life.

SA can also be used to extract and analyze subjective information from Web that includes social media comments, online reviews and other similar text sources. The analyzed data helps in calculating the public's sentiments or opinions toward certain product, people or ideas and reveal the contextual polarity of the information [2].

Two main approaches can be used for the sentiment analysis i.e. machine learning approach and lexicon-based approach. In machine learning approach, the data is classified by training a classifier on the labeled data [3]. In the lexicon-based approach, lexical items from dictionary are assigned positive and negative polarities. The lexicon is composed of a defined list of sentiment words along with their intensities and polarities [4]. The lexicon-based approach involves the calculation of sentiment from the semantic orientation of phrases and words that occur in a sentence [5]

Lexicon based approach is considered as a simple and reliable as compared to machine learning approach since it avoids the need to develop a labeled training set. Moreover, it is also difficult to ensure the correctness of labeled data in machine learning approach. To develop a sentiment lexicon, some researchers use dictionary or corpus-based approaches. Corpus based approaches involve the determining of the patterns of co-occurring of words to determine the sentiments of the words and phrases. The dictionary-based approach helps to compile the sentiment words and use the antonyms and synonyms in WordNet to determine the sentiments of the lexical items [6].

Depending on the nature of the data and choice of the users, the process of sentiment analysis can be performed on three different levels. These three levels of analyses are; 1) document level sentiment analysis, 2) aspect level sentiment analysis and 3) sentence level sentiment analysis [3]. In Sentence level sentiment analysis, each sentence is classified as positive, negative or neutral. Here the sentence is considered as a separate unit expressing a single opinion.

The aim of our research is to develop an Urdu sentiment lexicon marked on a five-component Likert scale ranging from -2 to +2. The developed lexicon comprehensively covers the four major word classes i.e. nouns, verbs, adverbs and adjectives. Conclusive guidelines are developed to annotate the data with different polarities. The study also aims to test the accuracy of the Urdu sentiment lexicon by using manual and algorithm-based approaches.

The current study is organized as follows: Section 2 labeled as literature review highlights some seminal

research works related to the topic, Section 3 highlights the methodology undertaken for the research. Section 4 covers the results of data analysis, Section 5 provides the discussion of the results, and Section 6 concludes the research and discusses the future dimensions.

2. Literature Survey

Due to unavailability of resources in other languages, sentiment analysis in multiple languages often involves transferring knowledge from one resource-rich language to other resource-poor languages [7]. Majority of multilingual sentiment analysis systems employ English lexical resources such as SentiWordNet. A popular approach towards SA is to use a machine translation system to translate texts from languages into English. The original text is translated into English, and then English SA resources such as SentiWordNet are employed [7]. However, translation systems pose various problems such as; sparseness and noise in the data [8]. Sometimes, translation system fails to translate essential parts of the original text which can possibly reduce the text's original sentiments [9].

SentiWordNet assigns WordNet synsets to three categories: positive, negative, and neutral by using numerical scores ranging from 0.0 to 1.0 to indicate the degree to which the terms included in the synset belong to the corresponding category. SentiWordNet is built by performing quantitative analysis of glosses for synsets [10]. One drawback of SentiWordNet is that it assigns polarity at the syntactic level but fails to assign polarities to phrases such as "getting angry" or "celebrate a party" which correspond to concepts found in the text to express positive or negative opinions [11].

Moreover, multilingual lexical resources specific to sentiment analysis are also developed. The NTCIR corpus of news articles in English, Chinese, and Japanese is made up of information on sentiment polarity and opinions for sports and political news data [12].

Different techniques can be used in the extraction and tagging of lexicons. The extraction of SentiUnits using shallow parsing techniques in order to create a sentiment lexicon can be considered as one of the most authentic method [13]. SentiUnits are expressions which carry sentiment information in the sentence.

Cambria et al. [11] proposed a SenticNet, lexical resource based on a multi-disciplinary approach to identify, interpret, and process sentiment in the Internet. SenticNet is more suited for a concept-level sentiment analysis and can also be utilized to evaluate texts based on common-sense reasoning tools that require large input. It employs a Sentic computing methodology, in particular, to evaluate texts at document or sentence level. It performs the task of building a collection of concepts, including common-sense concepts, supplied with positive or negative polarity labels. Unlike SentiWordNet, SenticNet does not assume a neutral polarity. It guarantees high accuracy in polarity detection with the availability of multilingual tools as well.

Many researchers use Semantics in creation of lexicon for performing SA. For the SA of twitter, a lexicon-based approach is presented called SentiCircles [14]. This approach considers the patterns of words that occur mutually according to different contexts, get their semantics and then update the sentiment lexicon accordingly by updating the pre-assigned polarity and strength of these patterns. Sentiment Knowledge is encoded into pre-trained word vectors for improving the performance of SA, where the proposed method is based on external sentiment lexicon and a convolution neural network.

Remus et al [15] worked with German inquirer, which is a German sentiment lexicon supplied with positive and negative labels. It was constructed using Google translate by translating words and terms into the German language. The words without any sentiment were removed from the German Inquirer. SEL is a Spanish emotion lexicon that contains 2036 words marked with the Probability Factor of Affective use (PFA) as the measure of their expression of basic emotions: joy, anger, fear, sadness, surprise, and disgust, on the scale of null, low, medium, or high. This lexicon was marked manually by 19 annotators who had to agree on a certain threshold for a label on the word to be included in the lexicon. Probability Factor of Affective use was developed by the authors of SEL to incorporate agreement between annotators in the decision-making process of labeling the sentiment on a word.

Mobarz et al. [16] created a sentiment Arabic lexical Semantic Database (SentiRDI) by using a dictionary-based approach. The database has many inflected forms, i.e., it is not lemma-based. Moreover, the authors reported insufficient quality and plan to try other alternatives.

Different researchers have also developed Urdu sentiment lexicons. An Urdu corpus, labeled with semantic role by using cross lingual projection, is developed [17]. Syed [18] proposed an innovative sentiment annotated lexicon for Urdu based on SentiUnits. Syed started by extracting SentiUnits i.e. positive and negative expressions, from a given Urdu text, using shallow parsing technique. Hashim and Khan [1] developed a sentiment analyzer based on Urdu Nouns and Adjectives for sentence level sentiment analysis. Hashim used Urdu news data from headlines by using a lexicon based on nouns and adjectives. Mukhtar and Khan [19] used a lexicon-based approach

for sentiment analysis of Urdu blogs, using a publicly available Urdu Sentiment Lexicon [20]. They included adjectives, nouns and negations; as well as verbs, intensifiers and context-dependent words. The developed Urdu sentiment analyzer applies rules, use lexicon and perform Urdu sentiment analysis by classifying sentences as positive, negative or neutral.

A lot of work has been done previously on the difficulties, strategies and the utilization of sentiment analysis of English language. The sentiment analysis techniques designed for English language cannot be utilized for Urdu language due to its different script, morphological and syntactic patterns. Because of the different structure of Urdu language, other languages SA cannot be employed for resolving the issues of Urdu language. Moreover, Urdu Sentiment Lexicons developed so far have covered the categories of nouns, verbs and adjectives while little or no attention has been paid on adverbs. However, this study stands out as it proposes comprehensive guidelines for developing an Urdu sentiment lexicon. This research also proposes a five component Likert scale ranging from -2 to +2 whereas the sentiment lexicons for Urdu language developed so far do not actually put forward such a comprehensive five level Likert scale. A maximum of three scale i.e. -1 to +1 Likert scale for Urdu sentiment Lexicon has been anticipated so far. Moreover, this developed lexicon comprehensively covers the four major word classes i.e. nouns, verbs, adverbs and adjectives.

3. Methodology

This section elucidates the research procedure, sampling technique and the aspects of data analysis tool and procedure.

The research is mixed method in nature and the research design is cross sectional. The methodological procedure is divided into two phases:

i. Development of Urdu Sentiment Lexicon
ii. Testing of Urdu Sentiment Lexicon

3.1. Development of Urdu Sentiment Lexicon

Manual tagging techniques such as dictionary-based methods were used to develop Urdu Sentiment Lexicon. For the manual tagging of the data, a comprehensive set of guidelines were also developed. The established guidelines provide an elaborative framework of tagging various word classes namely: nouns, verbs, adjectives and adverbs. Five point Likert scale comprising of -2, -1, 0, 1, 2 values was used to assign polarities to lexical items. The development of sentiment lexicon was further divided into two stages:

3.1.1. Data Extraction. At the first stage the lexical items to be tagged are extracted from a selected corpus [21] and around 35 M words were extracted. The sample undertaken for the development of Sentiment Lexicon included around 21556 words which mark the most frequently occurring words from the selected corpus which were already assigned with POS tags [21]. The extracted POS categories with their numbers are presented in Table 1 :

Table 1 POS categories and number of words in each category

POS Categories	Number of words
Nouns	15826
Verbs	3097
Adjectives	1986
Adverbs	646

3.1.2. Data Tagging. At the second stage, the extracted lexical items were tagged according to the developed guidelines by a team of linguists. Three authentic Urdu dictionaries [22] [23] [24] were also used to analyze the lexical items. The postulates of the formulated guidelines are as follows:

1. Assign higher polarities to the words that give clear sense of positivity or negativity while assign lower polarities to words that show rather vague sense. Neutral should be only those words that do not show any orientation.
2. Words that show some positive or negative sense without any context should be tagged according to their prior polarities. Moreover, words that depict strong positivity or negativity are assigned stronger polarities i.e. +2 or -2. For instance, the polarity of /دوست/ /friend/ /ḏo:sṯ/ is +2 and the polarity of /دشمن/ /enemy/ /ḏuʃmən/ is -2.
3. Words having multiple parts of speech tags whose polarities change according to the POS category are assigned respective polarity for each tag. For instance; the polarity of the word / محسن / /mohsın/ (Noun) name will be assigned a 0 polarity, while /محسن/ /friend/ /mohsın/ (Adjective) will be marked as +2.
4. The polarity of the words increase with the increase in the degrees of adjectives. For instance, the words /بد/ /bad/ /bəḏ/ and / بدترین/ /pathetic/ /bəḏṯəri:n/ will be assigned -1 and -2 polarities respectively. For adjectives ending at 'ی', orientation of the root word should be checked and assign the polarity accordingly. For instance the adjectives like; /قومی/

/national/ /kɔ:mi:/ and / /اندھیری /dark/ /ənd̪ʰe:ri:/ will be assigned 1 and -1 polarities respectively.

5. The singular words will have low polarity strength whereas; their plurals will have higher strength of polarity. Polarity increases with the increase in numbers. For instance, the singular noun /مشکل/ /difficulty/ /muʃkıl/ and its corresponding plural form /مشکلات/ /difficulties/ /muʃkıla:t̪/ /will be assigned -1 and -2 polarities respectively.
6. Nouns that represent ranks, titles or respect will be tagged with higher polarities. For example, /صاحب/ /Sir/ /sa:hıb/ and / حضرت /hazrat/ /həzrət̪/ will be assigned +2 polarity.
7. Words showing sense of certainty either positive or negative will be given higher polarities and those showing uncertainty will be assigned lower polarities. For instance, the adverbs /اچانک/ /suddenly/ /ətʃa:nək/ and /روزانہ/ /everyday /ro:za:na:/ will be assigned a polarity of -1 and 1 respectively since /اچانک/ represents a sense of uncertainty and /روزانہ/ represents a sense of certainty.
8. Verbs that convey some proper meaning as a word will be given higher polarities as compared to the words that themselves do not give proper meaning and need associating words with them. For instance /پڑھنا/ /study/ /pəɽhna/ and /رہتی/ /stays/ /ræhṭ̪:/ will be assigned +2 and +1 polarities respectively since /پڑھنا/ has proper meaning attached to it whereas /رہتی/ needs associating words to deliver its complete meaning.

3.2 Testing of Urdu Sentiment Lexicon

The developed Urdu Sentiment Lexicon was tested both manually and automatically:

3.2.1. Manual testing process. Manual testing process comprised of two steps:

1) 10% reference of the untagged original data was sampled automatically and marked manually by an expert linguist according to the guidelines
2) Another linguist assigned polarities to the original data set. Then the results of 10% tagged reference were compared with the polarity assigned original data to depict an inter annotator agreement between the two data sets. A threshold of 75% accuracy was established to check the quality of the data. The tagged source data found below the above-mentioned accuracy was sent back to the linguist for the further review.

3.2.2 Automatic testing process. For the purpose of analyzing the reliability of our lexicon, a baseline system was also developed and the accuracies were computed. The lexicon was tested using a subset of 500 sentences extracted from Roman Urdu sentiment dataset UCI machine learning repository [25] .These 500 roman Urdu sentences were then transliterated into Urdu before being processed. Following algorithm was used for automatically tagging the sentiment of a given input sentence S :

a) Remove special symbols and punctuations from the sentence S
b) Compute total positive words (tpw) in a sentence S
c) Compute total negative words (tnw) in a sentence S
d) Now assign sentiment to S using the following rules
   i. If tpw>tnw, assign positive sentiment to S
   ii. If tnw>tpw, assign negative sentiment to S
   iii. if tpw=tpw, assign neutral sentiment to S

After the application of the above algorithm, the accuracy of the lexicon was determined.

4. Results

This section elaborates the results obtained after the lexicon is manually and automatically tested.

A lexicon of around 21556 words was developed. The developed lexicon was manually tested and a 75% inter-annotator agreement was achieved. Around 25% data showed a mismatch of polarities due to the subjectivity of the annotator and the influence of the pragmatic stance during the data tagging process.

For the further validation of the lexicon, an automatic process was also used. A sample of 500 sentences (already assigned with a positive, negative or a neutral sentiment) from UCI corpus was taken. Then, the sentences were automatically tagged using our lexicon and the marking of lexicon and UCI corpus was compared. Through this process, an overall accuracy of 84.0% was achieved. Both manual and automatic testing results validated the lexicon for its validity and accurate nature.

5. Discussion

Since the lexicon was tagged on a five-point Likert scale which implies that word can be assigned a

sentiment of being positive, very positive, neutral, negative or very negative. The polarities of the lexical items can greatly be influenced by the annotators' subjectivity, opinion, pragmatics and contextual domain in which the word is occurring. For instance, words like; /حکومت/ /government/ /hʊku:mət̪/ can have different orienta tion. If the annotator is a supporter of the current government then he would mark it as +1 but if he is not a supporter so he will mark it as -1.
Also, if the adjective / تیز / /fast/ /t̪e:z/ is considered then the word would carry a positive sentiment but, in the sentence,

/وہ بہت تیز ہے/
vo: bo:hət̪ t̪e:z hæ:
she very cunning is
She is very cunning,

The same word would carry a -2 sentiment due to the negative sense enacted by the word. Due to all these subjectivity problems, it was instructed to the linguists to not to think the contexts of the words rather mark them in their dictionary meaning that's why three online dictionaries were used for the analysis.

In addition to this, there were certain words whom meaning varied by the addition of diacritics. For instance the word سونا can either be سُونا/su:nɑ:/ desolate/ or سونا /sleep/ /so:nɑ:/. Also, the word classes vary due to the presence of diacritics where سُونا is an adjective and سونا is a verb. That is why, we used POS tagging and separated the words into their different classes to mark them easily and get the inter- annotator accuracy.

The data was also validated using a corpus-based approach and the polarity of the whole sentence was determined by the sum of the polarities of individual lexical items in the sentence. The testing also posed the good accuracy of 84%. However, there were certain sentences where the polarities assigned by the algorithm of the developed Urdu Lexicon and already assigned polarities of the UCI corpus showed a discrepancy. For instance,

/اللہ اللہ کر کے آئے/
/Thank God they came/
/əlla:h əlla:h kərke: a:e:/

The above sentence had a negative sentiment attached in the UCI corpus while the sentiment assigned by the developed lexicon using automatic means was positive. Since, we being the Muslims attach positive sentiment to اللہ i.e +2 and same was assigned in the Sentiment lexicon. So, a mismatch was found. The reason might be that they marked the sentence in the contextual sense but we did not consider the context while marking the lexical items.

6. Conclusion and Future Directions

The sentiment lexicons hold a great importance in defining the semantics of particular lexical items. The current study defines the development of an Urdu sentiment lexicon. The current research can be regarded as authentic since it consists of a wide repository of lexical items (nouns, adjectives, verbs and adverbs) marked on -2 to +2 Likert scale. Also, the manual testing results yielded an accuracy of 75% and the accuracy was further validated by the automatic testing method where the percentage accuracy attained was 84%. The findings of this lexicon can further be applied on bigrams including collocations and phrasal verbs and the repository of the words can be enhanced. However, this data can also be further validated using machine learning techniques in future.

The developed lexicon can also be used to create a business intelligence system and can provide aid in the sentiment analysis of a particular corpus, which can include online reviews, feedbacks and twitter comments. The opinion mining of news data can also be executed through it. Moreover, this lexicon can be integrated in any programming language to develop an automated sentiment analysis system.

Since the lexicon was automatically tested and an overall accuracy of 84% was achieved, this number could be further improved by modification of the algorithm and the percentage accuracy of the lexicon can be taken to 95% and above.

Different dimensions can further be explored e.g, polarity of modifiers, negations, pronouns and lexically borrowed words from English can be studied. The population of adjectives in the lexicon can be increased to further improve the lexicon since the adjectives hold a great stance in defining a sentiment.

7. References

[1] F. Hashim and M. A. Khan, "Sentence Level Sentiment Analysis using Urdu Nouns," in Conference on Languauge and Technology, Lahore, Pakistan, 2016.

[2] R. Eskander and O. Rambow, "SLSA : A Sentiment Lexicon for Standard Arabic," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.

[3] B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis," in Mining Text Data, Boston, Springer, 2012.

[4] P. Gonçalves, M. Araújo, F. Benevenuto and M. Cha, "Comparing and combining sentiment analysis methods," in Proceedings of the first ACM conference on

Online social networks, Boston, Massachusetts, USA, 2013.

[5] M. Taboada, J. Brooke, M. Tofiloski and K. Voll, "Lexicon-Based Methods for Sentiment Analysis," Computational Linguistics, pp. 267-307, 2011.

[6] X. Ding, B. Liu and P. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining," in 15th ACM SIGKDD international conference on knowledge discovery and data mining, 2008.

[7] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in IEEE 24th international data engineering workshop, 2008, 008.

[8] Balahur and M. Turchi, "Multilingual Sentiment Analysis using Machine Translation?," in Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis, 2012.

[9] M. Bautin, L. Vijayarenu and S. Skiena, "International sentiment analysis for news and blogs," in ICWSM, 2008.

[10] . V. Singh, R. Piryani, A. Uddin and P. Waila, "Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches," in 5th international conference on knowledge and smart technology (KST). IEEE, 2013.

[11] E. Cambria, R. Speer, C. Havasi and A. Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining," AAAI fall symposium: commonsense knowledge, p. 02, 2010.

[12] Y. Seki, D. K. Evans and L.-W. Ku, "Overview of Multilingual Opinion Analysis Task at NTCIR-7," in Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering, and cross-lingual information access, 2008.

[13] D. Jayraj and S. Andhariya, "Sentiment analysis approach to adapt a shallow parsing based sentiment lexicon," in International Conference on Innovations in Information, Embedded and Communication Systems, 2015.

[14] H. Saif, Y. He, M. Fernandez and H. Alani, "Contextual Semantics for Sentiment Analysis of Twitter," Inf Process Manage , pp. 5-19, 2016.

[15] R. Remus, U. Quasthoff and G. Heyer, "SentiWS-a publicly available German-language resource for sentiment Analysis," in LREC, 2010.

[16] H. Mobarz, M. 7Rashwan and I. AbdelRahman, "Generating lexical Resources for Opinion Mining in Arabic Language Automatically," in The Eleventh Conference on Language Engineering (ESOLE), Cairo-Egypt., 2011.

[17] S. Mukund, R. Srihari and E. Peterson, "An information–extraction system for Urdu—a resource poor language," ACM Transactions on Asian Language Information Processing, pp. 1-43, 2010.

[18] S. Afraz Z, M. Aslam and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in Mexican International Conference on Artificial Intelligence, Mexico, 2010.

[19] N. Mukhtar and M. A. Khan, "Effective lexicon based approach for Urdu sentiment analysis," Artificial Intellegence Review, 2019.

[20] Athar, "Chaoticity," 14 June 2012. [Online]. Available: https://chaoticity.com/urdu-sentiment-lexicon/

[21] M. Farooq and B. Mumtaz, "Urdu Phonological Rules in Connected Speech," in Conference on Language and Technology, Lahore, Pakistan, 2016.

[22] "Urdu Lughat," 2007. [Online]. Available: http://udb.gov.pk/.

[23] "Online Urdu Dictionary," 2002. [Online]. Available: http://www.cle.org.pk/oud.

[24] "Urdu Lughat," 2013. [Online]. Available: http://urdulughat.info/.

[25] Z. Sharf and S. Rehman, "UCI Machine Learning repository," IJCSNS International Journal of Computer Science and Network Security, vol. 18, no. 1, January 2018.

[26] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Hawalah, A. Gelbukh and Q. Zhou, "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognitive Computation, pp. 757-771, 2016.

[27] M. Humayoun, H. Hammarström and A. Ranta, "Urdu morphology, orthography and lexicon extraction," in 2nd Workshop on computational approaches to Arabic script-based languages, Stanford, 2007.

[28] M. Ijaz and S. Hussain, "Corpus Based Urdu Lexicon Development," in Conference on language technology (CLT 2007), 2007.

[29] Muaz, A. Ali and S. Hussain, "Analysis and Development of Urdu POS Tagged Corpus," in 7th Workshop on Asian language resources, Suntec, Singapore , 2009.

[30] R. Tatman, "Sentiment Lexicons for 81 Languages," 2017. [Online]. Available: https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages.

[31] U. Mirchev and M. Last, "Multi-document summarization by extended graph text representation and importance refinement," Innov Doc Summ Tech Revolut Knowl Underst Revolut Knowl Underst, 2014.