Lexicon-based Sentiment Analysis for Urdu Language

Zia Ul Rehman
Department of Computer Science & IT
University of Sargodha, Lahore Campus
Lahore, Pakistan
ziaulrehman@uoslahore.edu.pk

Imran Sarwar Bajwa
Department of Computer Science & IT
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
imran.sarwar@iub.edu.pk

***Abstract*— Social media has recently become as a powerful weapon people use for online discourse, creating content, share it and network with other individuals at a phenomenal frequency. With social media and user-generated content exploding the web/ blogs/ social networking forums, vendors/ critiques/ socialist and influential individuals got enthusiastic to mine this substantial data set for obvious meaning, but they soon discovered a novel challenge: to know that someone is talking about a particular topic/service/brand or social event is far less important in comparison to know how they are feeling and conversing about it. This is known as sentiment analysis or opinion mining. Numbers divulge that people are extensively using social media, expressing their positive opinions or negative apprehensions online. As an aftermath the concept of sentiment analysis/ opinion mining is broadly being acknowledged and employed by society as a whole to enhance their business/products/services or just to assess the overall prevailing environment. Acknowledged work is being done in this area converging towards exploring sentiment analysis, its definite requirement in this era, different frameworks for sentiment analysis and their comparison with other previously proposed techniques, but unfortunately Urdu language is not considered comprehensively in this context. As Urdu is one of the prevalent languages, this paper aims at creating an application for sentiment analysis of Urdu comments on various websites. Elaborated system architecture is discussed in detail with techniques employed; experimentation procedure and proven results of 66% accuracy are also deliberated. The F-measure achieved by this proposed system is 0.73. Challenges faced in sentiment analysis with respect to this neglected language are also highlighted for future considerations.**

***Index Terms*—Sentiment Analysis, Opinion Mining, Artificial Intelligence, Natural Language Processing, Urdu text processing.**

I. Introduction

Natural Language Processing (NLP) is an emerging field of artificial intelligence that deals with assessing, understanding, analyzing and generating the languages that humans practice naturally to interface with machines. NLP applications are achieving approval as human-machine interactions are increasing. These applications can be categorized into two distinct aspect. The computational dimension focuses on machine translators, grammar checkers and spell checkers for natural languages used by humans for interacting. On the other hand understanding and manipulating conceptual/psychological aspects of natural languages is altogether different dimension of NLP applications. Conceptual understanding encompasses interpretational side also. Story/poetry generation, intelligent information gathering/retrieval, syntactic analysis, pragmatics and sentiment analysis are all sub-fields to this dimension. In this paper we emphasis on the evolving research area of sentiment analysis.

In principle, sentiment analysis is defining the emotional tone behind a sequence of words, used for understanding attitudes, emotions and opinions conveyed within an online forum. The previous research in this context predominately address languages such as English, Arabic, Italian, Chinese and some other widely used languages. The techniques proposed here are not suitable for dialects of South Asia as they have altogether different morphology, script and grammatical notions, for instance Urdu is one predominant language used by 68.5 million (http://www.ethnologue.com) but no considerable work has been proposed. Noticing this deficiency, in this research paper we propose a sentiment analyzer for Urdu language as it predominant language of Asia. Secondly with increased usage of blogs, webs, users feel comfortable to express sentiments in their native language, consequently intensified usage of Urdu language has been witnessed in few years.

The paper is structured in this manner: Section II proceeds by a brief overview on sentiment analysis/opinion mining. Section III initiates literature review by deliberating requirement for Urdu sentiment analyzer, detail summary on previous work in the respective context is also considered. In section IV proposed methodology for Urdu sentiment analyzer is discoursed in an elaborative method. Algorithm, results of proposed technique and deep analysis is demonstrated in section V. Finally section VI concludes the research, discussing limitations and exploring new research dimensions for future.

II. Overview of Sentiment/Opinion Mining

Sentiment analysis, broadly termed as opinion mining is a challenging discipline which aims at analyzing people sentiments, opinions, assessments, evaluations, attitude, behaviors, appraisals, feelings and emotions concentrated towards objects such as establishments, services, products,

978-1-5090-2000-3/16/$31.00 ©2016 IEEE

individuals, concerns and events [1]. Sentiment analysis predominantly emphasizes on visions which express/infer affirmative or undesirable sentiments. Opinions are key drivers to human activities; most significantly in decision making process other people's point of view is prime influencer. Emergent social media (blogs, forum discussion, comments, and postings) empowers individual of distinct creeds to conveniently express their opinions regarding any entity thus effecting other people decision.

These all factors determine the need of sentiment analysis application. These applications vary widely in domains such as consumer services/products, healthcare, social events and many more [2]. Sentiment analysis has been handled as natural language processing (NLP) task and executed at three levels; Document/Snippet level, Sentence level and aspect level [3][4]. Moreover we can perform sentiment analysis utilizing two techniques machine learning or by constructing lexicons [5]. Both techniques have their weightage in terms of benefits and short-comings.

III. Literature Review of Urdu Language

Urdu is one of the most pre dominant language of Asia and is member of an Indo-Aryan language. It is national language of Pakistan and is one of the scheduled languages in India. Urdu is a complex language as it consists of compound characters which mean two or more characters combined to form a complex shaped character. Moreover the word changes its meaning depending on its position in the sentence [6].Urdu shares script similarities with Arabic and Persian, morphological similarities with Hindi but even then it has individual existence as its computational linguistics are altogether different. This distinct language is lacking even basic NLP tools for sentiment analyzer [7]. The literature survey initiates with exploring research paper showing that prevailing techniques and approaches for other languages are not capable enough to handle Urdu text. Then literature survey proceeds with existing research work carried out for sentiment analysis of Urdu language. The considerable work published over the time period of 2010 to 2016 is considered.

Kashif Riaz [8] in his research paper presents the challenges faced in writing Urdu stemmer/ morphological analyzer. The diverse nature of Urdu and lack of machine readable resources are focal hindrance. Morphology works through inner structure and construction of words, and is vital component for informational retrieval applications. The author claims that no current stemmer can be employed, discussing issues critical to creating stemmer specifically for Urdu language. S. Hussain [9] discourse about the importance of derivational and inflection morphology and how it is employed in various forms of applications. Author then claims that no reported work has been done on Urdu stemming and how current stemmers are unbefitting for the requirements of Urdu language structure. After this a new proposed Urdu stemmer labelled as "Assas-Band" is elaborated. This novel stemmer provides accuracy up to 91 % replacing old technique of conventional lexical lookup by proficient exceptional lists. The results are presented and proved.

In 2010 Afraz Z. Syed [7] in his work proposes an innovative sentiment-annotated lexicon for Urdu language. The approach works by extracting SentiUnits (negative and positive expressions) from given Urdu text employing shallow parsing. The major contribution is development of lexicon highlighting linguistic and technical aspects of Urdu language. The lexicon renders satisfactory results. In the same year Mukund [10] develops a classifier which can distinguish between subjective and objective sentences for the Urdu language. The author utilizes VSM (Vector Space Model) and SVM (Support Vector Machine) to create learning techniques. Experiments show that VSM approach is much more efficient then SVM approach. Author contributions are acknowledged as generating data sets for Urdu language which has no annotated data was difficult task. Mukhand [11] also developed an information extraction system for Urdu language. The basic aim of this research was to provide NLP Infrastructure with entity tagging.

In 2011 Afraz Z. Syed [12] in his next research inspects phrase-level negation on the sentiment analysis of Urdu reviews. This approach uses subjective text SentiUnits. The proposed analyzer deliberates each sentence one by one, mines SentiUnits and computes the polarity of the sentence. Despite being innovator effort results of this approach are positive. In the same year Mukhand [13] implemented sentiment analysis utilizing linear kernal (LK) and sequence Kernel (SK) methods with satisfactory results. In proceeding year of 2012 Afraz Z. Syed [14] utilizing his previous research on SentiUnits created a sentiment-annotated lexicon for Urdu words. In this work he clearly marked orientations (positive/negative) for the corpus of movie and electronic devices reviews. In 2013 Hammad Afzal [15] carried out new research for spatial analysis of bi-lingual tweets for a social event. The proposed application showed encouraging results but with loop hole as multiple words can be used to classify one distinctive event.

In 2015 Muhammad [19] compared and contrasted three diverge techniques for sentiment analysis of Roman-Urdu comments from a blog. The author tested models and his results indicated that Naïve Bayesian technique outperformed Decision Tree and KNN techniques in all the performance metrics. In 2016 S.Abbass [16] took an innovative step in finding salience in Urdu news utilizing Heuristic technique. His focus was to analyze political news, calculating overall polarity.

IV. Proposed Methodology

The proposed system works on associating polarity to given Urdu sentence or piece of Urdu text. In our system we have specifically focused on comments/opinions posted on a certain topic by reviewers of news website (blog.jang.com.pk) and then allocating polarity (positive, negative or neutral). System architecture is elaborated in next section, before that we had two challenges to accomplish which are discussed below:

A. Lexicon Improved/ Reformed

This system utilizes Urdu lexicon available publically at https://github.com/ which lists 2607 positive and 4728 negative sentiment words for Urdu. Table I shows a sample of the Urdu lexicon used in proposed methodology.

TABLE I. A SAMPLE OF URDU SENTIMENT LEXICON

Word	Label	Word	Label
مہربان	1	نااہل	-1
شریف	1	شرمناک	-1
عمدہ	1	برا	-1
مفید	1	اچھوت	-1
امن	1	اوباش	-1

This lexicon is originally based on technique for mining customer reviews [17]. English Dictionary based lexicon is translated into Urdu language sentiments. We modified this lexicon and removed irrelevant words included for Urdu language. Comprehensive manual review is performed on this dictionary adding/removing word for enhancing polarity capacity. In total this lexicon has 7335 items. Positive words are assigned polarity of 1 and negative words are presented by -1.

B. Corpus Construction:

A huge amount of text is required to construct the corpus which is used for lexicon analysis. Electronically accessible resources are the suitable method for gathering of text but unfortunately for Urdu it's not readily available. Secondly most of the data is in graphics/image form which is not extractable [18]. In our proposed system, we have extracted text from Urdu news websites such as bbcurdu.com, dawnnews.tv and user opinion at blog.jang.com.pk about a specific topic. There is no publically available corpus for Urdu sentiments.

C. System Architecture

The system architecture of Urdu sentiment analysis by implementing a sentiment classification for Urdu opinion is elaborated in Fig. 1. At first stage comment undergoes the pre-processing stage. Tokens are generated from the sentence then passed to polarity identification stage. Polarity of each word is assigned equating with sentiment lexicon.

Polarities are assigned as: Positive=1, Negative=-1, Neutral=0. Once individual polarities are calculated the overall polarity of the sentence is determined by weighing negative or positive indication. For instance if a particular sentence has two positive words and one negative word, overall polarity would be calculated as + 1(+2-1), hence declaring it as positive comment. The output shows whether the comment/opinion of the viewer had a positive, negative or neutral sentiment towards the news.

V. EXPERIMENTATION , RESULTS AND ANALYSIS

This segment is aimed at clarifying practical implementation, results and analysis of the proposed technique. It will initiate with explanation of pseudocode, discussing the working for all scenario, then it will proceed with practical work and it will conclude with results. Analysis on the results provide comprehensive efficacy of the Urdu Sentiment analysis thus making overall performance perceptible.

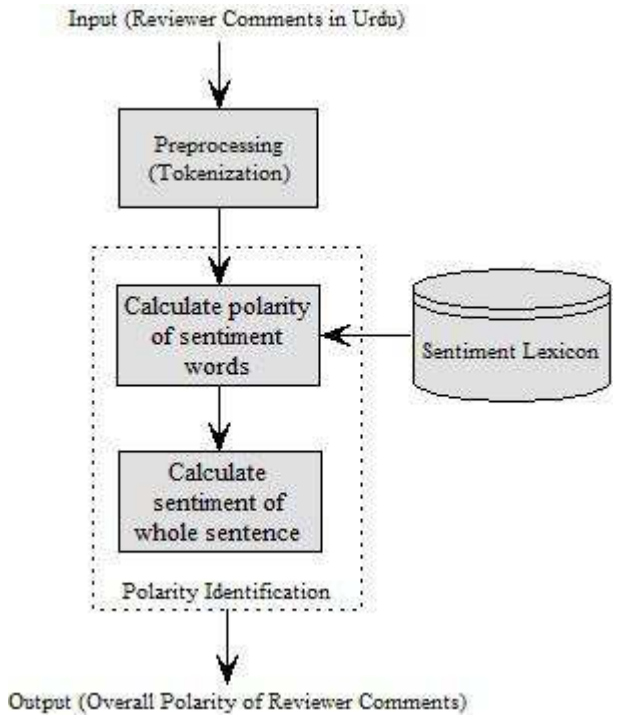

Fig. 1. System Architecture

A. Pseudocode and Evaluation Framework

The pseudocode for the lexicon-based sentiment analysis for Urdu text is given below in Frame 1. The polarity of a comment is purely determined by tokens of the comment and the sentiment lexicon. In the current work, the polarity of a given comment is calculated as explained in the procedure *PredictSentiment* that deals with three situations in predicting the nature of sentiment: positive, negative and neutral.

- The viewer of the news/social event or blogs can have positive feeling towards its. For instance, اسلام محبت اور امن کا درس دیتا ہے (Islam teaches love and peace). In this text "محبت" and "امن" both are positive words and polarity score is +2 so overall sentiments are positive.
- Negative sentiment will show polarity score in negative numbers. Considering this comment کرکٹ بورڈ کے ارب اختیار نااہل ہیں (Cricket board authorities are incompetent), "نااہل" is a negative opinion. Polarity calculation will show -1 thus rendering it as a negative opinion.
- Neutral comments are also common. The viewer dislikes or likes certain dimension of the topic but there are an equal number of positive and negative opinions. Let consider this remark ہماری ٹیم جیت بھی سکتی ہے اور ہار بھی (our team can win and lose), "جیت" is positive word (+1) and "ہار" is negative (-1) so overall polarity score will be zero. Consequently this sentence will be marked as neutral sentiment.

FRAME 1. PSEUDOCODE FOR DETERMINING POLARITY

Given an Urdu text as input, the system proceeds as follows:

Procedure **PredictSentiment()**

```
1. begin
2.        sentiment;
3.        rank = 0;
4.        For each opinion word in lexicon
5.           w = polarity of word in lexicon;
6.          /*Positive = 1, Negative = -1*/
7.           if(w=positive)
8.                  rank+=1;
9.           else if (w=negative)
10.                 rank+=-1;
11.       if (rank<0) sentiment = negative;
12.       else if (rank>0) sentiment =positive;
13.       else sentiment = neutral;
14. end
```

Evaluation framework is used to calculate the overall performance and quality of the proposed system. Five metrics are calculated; accuracy, recall, precision, F-score and Error-rate. These all metrics are defined using confusion matrix. In consideration with proposed system formula for these five metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Score = Harmonic\ mean\ of\ precision\ and\ recall$$

$$Error - rate = \frac{FP + FN}{TP + FP + TN + FN}$$

Where TN, TP, FN and FP are the numbers of true negatives, true positives, false negatives and false positives respectively.

B. Experiments and Results

In this set of experiments, the Urdu sentiment lexicon was employed to define the polarity of comments on Urdu news website. In Experiment one twenty four comments were subjectively classified using proposed sentiment analysis framework. During preprocessing tokens were generated for the input sentence. These tokens were equated with Urdu lexicon to determine the polarity of the words. The results are shown in Table II, we have calculated all five metrics with above proposed methodology of Urdu sentiment analysis.

The results of lexicon based approach are satisfactory. We can improve results if we handle certain complications associated with Urdu text. Technically we face many glitches as we proceed with sentiment analysis of Urdu language such as unavailability of electronic data and lexicon for Urdu language. Word segmentation also hinders competent analysis and it is pertinent to Urdu language.

TABLE II. RESULTS OF LEXICON-BASED APPROACH

Accuracy	0.66
Recall	0.79
Precision	0.69
F-Score	0.73
Error-Rate	0.35

For instance شرمناک and شرم ناک are two same words with same meaning (in English its meaning is: embarrassing), lexicon assigns شرمناک as a word with negative polarity but if reviewer writes شرم ناک no polarity would be assigned rendering wrong results.

Diversity/flexibility in Urdu language also effect results negatively. People express their sentiments in English words but write in Urdu for example ہماری قوم ہر کام میں ٹیلنٹ رکھتی ہے and آپ کا بلاگ بہت لائک کیا گیا. In these examples above ٹیلنٹ and لائک are the English words "talent" and "like" respectively which real meanings in Urdu are قابلیت and پسند. These are some examples from Urdu comments where lexicon will not detect the sentiment words as they are written in English thus assigning undetectable polarity. Accuracy of results is compromised in this scenario.

VI. Conclusion & Future Work

This paper has presented a novel framework for sentiment analysis in Urdu comments. The lexicon based architecture works by assigning polarities to the tokens generated by Urdu sentence. The lexicon has 7335 entries; 2607 negative and 4728 positive. The overall polarity of sentence is summation of all respective terms weight. Experiment on the data set of one hundred and twenty four Urdu comments from various Urdu website is performed to check the effectiveness of proposed framework. The architecture shows overall efficiency of 66 %.

Sentiment Analysis in Urdu is a challenging field for innovative research aspects. Even existing research in this context requires enhancement for improved results. Constructing extended lexicon with bulk electronic Urdu text, managing various dialects and handling several technical and linguistic glitches can be focused in future.

References

[1] Pang, B., Lee, L.: "Opinion Mining and Sentiment Analysis", in "Foundations and Trends in Information 956 Retrieval", Volume 2, Issue 1-2, January 2008, pp. 1-135

[2] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89.

[3] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing,vol.10, 2002, pp. 79-86

[5] S.M.Vohra and J.B.Teraiya. “A comparative Study of Sentiment Analysis Techniques”, Journal JIKRCE, vol. 2, no. 2, pp. 313-317, (2013)

[6] Vivek Kumar Singh , “A Survey of Sentiment analysis in Urdu”, INDJSRT 2015

[7] Syed, Afraz Z., Muhammad Aslam, and Ana Maria Martinez-Enriquez. "Lexicon based sentiment analysis of Urdu text using SentiUnits." Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2010. 32-43.

[8] Riaz, Kashif. "Challenges in Urdu Stemming (A Progress Report)." BCS IRSG Symposium: Future Directions in Information Access. 2007.

[9] Akram, Qurat-ul-Ain, Asma Naseer, and Sarmad Hussain. "Assas-Band, an affix-exception-list based Urdu stemmer." Proceedings of the 7th Workshop on Asian Language Resources. Association for Computational Linguistics, 2009.

[10] Mukund S &Srihari RK (2010). A vector space model for subjectivity classification in Urdu aided by co-training. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 860-868.

[11] Mukund S, Srihari R & Peterson E (2010). An Information-Extraction System for Urdu - A Resource Poor Language. ACM Transactions on Asian Language Information Processing (TALIP), 9(4) 15.

[12] Syed, Afraz Zahra, Muhammad Aslam, and Ana Maria Martinez-Enriquez. "Sentiment analysis of urdu language: handling phrase-level negation."Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2011. 382-393.

[13] Mukund, Smruthi, Debanjan Ghosh, and Rohini K. Srihari. "Using sequence kernels to identify opinion entities in Urdu." Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011.

[14] Syed, Afraz Z., Muhammad Aslam, and Ana Maria Martinez-Enriquez. "Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text." Artificial Intelligence Review 41.4 (2014): 535-561.

[15] Javed, Iqra, and Hammad Afzal. "Opinion analysis of Bi-lingual Event Data from Social Networks." ESSEM@ AI* IA. 2013.

[16] Ali, S. Abbas, et al. "Salience Analysis of NEWS Corpus using Heuristic Approach in Urdu Language." International Journal of Computer Science and Network Security (IJCSNS) 16.4 (2016): 28.

[17] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[18] Ijaz, Madiha, and Sarmad Hussain. "Corpus based Urdu lexicon development." the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan. Vol. 73. 2007.

[19] Bilal, Muhammad, et al. "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques."Journal of King Saud University-Computer and Information Sciences (2015).