Salience Analysis of NEWS Corpus using Heuristic Approach in Urdu Language

***S. Abbas Ali, M. Daniyal Noor, Munir Ahmed Javed, M. Mohsin Aslam, Omer Ahmed Khan, Noor us Sahar Zubair**

Department of Computer & Information Systems Engineering, N.E.D. University, Pakistan.

Summary
This research presents an innovative step in finding important Entity and Salience in Urdu News corpus by finding Polarity (degree of positivity or negativity) of News for any particular Entity using Heuristic Approach. This research focuses on analyzing the political News Corpus for finding Important Entities, Saliences in the Urdu language and calculating the overall Polarity of the News for each particular Entity.Online Urdu newspapers are used for processing different steps involved in this work. The steps involved in determining salience are Parts Of Speech tagging, Finding important entity, saliences in the corpus, assignment and calculation of Polarities for particular Entity. It is an initial step in Heuristic based Salience Analysis of Urdu News Corpus. The Accuracy achieved by this system is 84.5%.

Key words:
Natural Language Processing, Salience Analysis, Opinion Mining, Polarity Computation, Sentiment Analysis, Urdu News Processing,

1. Introduction

Salience Analysis is a recent research area in the field of Natural Language Processing (NLP). Immense work has been done in these fields for languages like English and German. But languages like Arabic and Urdu due to their complex sentence structure are not under much consideration for NLP. Researches have been done and projects have been made on some parts of this Research like POS Tagging but no such complete solution is provided for Salience Analysis or Opinion Mining especially for Analyzing News.

In the domain of opinion mining much work has been done but their methodology and objective are different from this research. Sentiment analysis is well explored for English language and is highly domain and language specific, Bo Pang et al., 2008 presented the challenges and different aspects of opinion mining and sentiment analysis [1]. Mukund et al., 2011 has described an approach to identify opinion entities using a combination of kernels in Urdu [2]. Afraz Z. Syed et al., 2010 used sentiment-annotated lexicon based approach for sentiment analysis of Urdu text [3] [4]. A survey of opinion mining and analysis of blogs, Review sites and micro blogging for English language is presented by Arti Buche et al., 2013 [5]. Vijiya Lakshmi et al., 2015 described the sequence labeling approach to opinion prediction and built the models for classifying opinion about a particular product review for English language [6]. The challenges and Complexities in processing Urdu language is discussed in detail by Waqas Anwar et al., 2006 [7]. Jasmine Bhaskar et al., 2014 described the sentiment classification of product reviews by considering objective words and intensifiers using SentiWordNet [8].Calculation of overall polarity using more linguistic and structural information and achieving the goal of Salience Analysis using more resources using Machine Learning is presented in [24]. This research paper will discuss the complete methodology for Analyzing Political News Corpus for finding Important Entities and Saliences in the Urdu News corpus and calculating the overall Polarity of the News for each particular Entity. Some major modules of Salience Analysis includes: POS Tagging, Finding Entities, and Finding Saliences, Associating each Salience with an Entity, Assigning Polarities to each Salience and Calculating Overall Polarity of News for any particular Entity. Rest of paper is organized as follows. In Section II, Corpus and words lists collection are defined. Section III describes research methodology and techniques including: parts of speech tagging, finding important entity, saliences finding and calculation of polarities for particular Entity. Experimental results are presented in section IV. Finally, conclusions are drawn in section V.

II. Corpus and words lists collection

The reason of Urdu language being a vast research area is its growing content available from different sources now a days [9].For this research work, Urdu News Corpus is collected for processing and different Words Lists that will be required during different steps in this work for achieving the goal of Salience Analysis. Issues regarding Urdu corpus construction are discussed by Kashif Riaz et al., 2002 [10].This research focuses Political News for analysis so the corpus are collected from well-known News Websites like DAWN News, ARY News, Nawai Waqt and BBC Urdu.

Manuscript received April 5, 2016
Manuscript revised April 20, 2016

Table 1: Sources of Corpus

Source	Number of Corpus
Nawai Waqt	50
DAWN News	50
ARY News	50
BBC Urdu	50

Table 2: Type of Corpus Used

Type of Corpus	Number of Corpus
Raw Corpus	200
POS Tagged Corpus	200
Manually Tagged Corpus	100
Raw Corpus	200

Several works have been done for analysis and development of Urdu POS Tagged corpus [11]. Following are the number of corpuses that are used for analyzing the overall behavior. The whole process of Salience Analysis requires Polarity Word Lists for calculating overall polarity of the News for any particular Entity. Following is the type and number of Words collected by Manually Tagging Words by software developed as part of this research work.

Table 3: Number of Manually Tagged Words

Type of Words	Number of Words
Entity	2951
-10 Polarity	355
-5 Polarity	1137
0 Polarity	154
+5 Polarity	853
+10 Polarity	111

Table 4: Existing Polarity Words Lists

Polarity of Words Lists	Number of Words
Positive	2617
Negative	4754

Following are the type and numbers of words that are taken from publically available Urdu Sentiment Lexicon [26].For collecting Corpus following standards are followed:

a) UTF-8 encoding is used to store the corpus in text files.
b) An Excel sheet is maintained to store the following data against each corpus.
- Complete URL of source of corpus.
- Date of that News.
- News Caption.
- Expected outcome of the system.
- Domain of the News.

III. Research Methodology and Techniques

After collecting corpus the first step is to normalize the text but as the text is taken from news so there is no need to normalize text for further processing. After that there is a need to cater Tokenization problems for breaking the corpus into chunks. There are several challenges like Joiner, Non-Joiner and Space insertion and Omission for tokenizing Urdu text [12] [13]. Word Segmentation with handling of Space Insertion and Omission for Urdu script is discussed by Gurpreet Singh Lehal [14] [15]. Then POS tagging will be done and to find those Entities on which the whole News depends, in the next step Saliences will be found i.e. those words that defines the polarity of the statement. Then polarities will be assigned to all the Saliences and finally calculation of the overall polarity specifically for one or more Entities recognition [23]. Fig. 1. illustrates the high level flow of different processes involved in this research. The following terminologies are used in this paper.

Corpus: The text of the whole News except the heading of the News.

Entity: The Person or any Object either Noun (NN) or Proper Noun (PN) discussed in the given News Corpus. A single News Corpus can be about one or more Entities.

Polarity: Degree of Positivity or Negativity of the word in general and not in the particular context of the corpus.

Salience: Any word that has some Polarity and is associated with one or more Entity in the corpus.

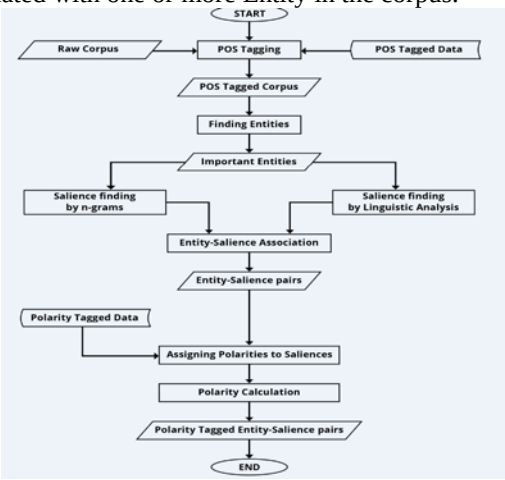

Fig. 1. Flow diagram of system processing

A) POS tagging

For POS Tagging, a pre developed Statistical Based POS Tagger [16] [17] from Center of Language Engineering (CLE), University of Engineering and Technology Lahore (UET) is used in this research. Brill's rule-based Tagging is discussed by Beáta Megyesi, 1992 [18], and a detailed comparison of Tagging techniques is presented by Hassan Sajjad et al., 2009 [19]. Detailed reference of Statistical

POS Tagger for Urdu is available in [27].In this work only few Parts of Speech has significance which are: Noun (NN), Proper Noun (PN), Adjective (ADJ), Verb (VB) Adverbs (ADV) and Negations (NEG). Entity can either be PN or NN, and Salience can be ADJ, VB or NN. Development of complex Tag sets for Urdu is defined in [20] and [21].

B) Entity Finding

The process of finding important entities is started to identify the main entity/entities of each individual news so that entity can be related to any attribute or quality or deed that can be classified as positive or negative. This process is to find the most relevant Entities that are discussed in the News and it is analyzed that it has very less chance that their number will be greater than three.

The criteria for the selection of important entities that can be selected as the object of the News, following major parameters were examined:

- That Entity should be a Proper Noun (or Noun in some cases).
- That Entity should be the most repetitive one among all others in the corpus.
- The Entities in Heading of News will be given more priority for selection.
- Usually the first Entity in the sentence is more important than the succeeding ones.
- Position of the Entity in the Sentence is important for getting contextual information.
- The Implementation Brill Algorithm in POS Tagger tags Noun (NN) to those words that are either NN or the word is not present in the POS Tagged corpus or dictionary, so to identify the difference in actual NN and ambiguous NN, priority has given to those that are actually present in dictionary and POS Tagged corpus.

All these features for the sake of normalization are given weightage out of 100 and then to calculate overall weight some percentages of that weight is taken after analyzing the behavior of each feature by taking different set of percentages. These percentages are given in the table below:

Table 5: Weightage of Major Features

Feature	% of the Weight given
Parts Of Speech	20%
Occurrence in Heading	30%
Total Occurrence in Corpus	30%
Occurrence/Position in First Sentence	10%
Occurrence/Position in Other Sentences	10%
Ambiguity Correction	25%

These are the basic outlines of our logic behind the Selection of important entity module. The detailed structure is defined below:

1) The above mentioned parameters are calculated by using the following criteria:
   a. First the file is broken into chunks and word tag lists are formed. Two consecutive Proper Noun (PN) are considered to be a single PN. Then this word tag list is transformed into sentences wise lists in which every Noun (NN), Proper Noun (PN) and their positions are stored.
   b. Distinct Entities have found so that there is no ambiguity between different versions of the same entity. For example if somewhere is written (Imran Khan) and at the other place only Imran (First name) is written then they would be considered to be the same.
2) To assign POS weight count of PN and NN tag for each Entity is counted if the Entity is tagged as PN more times, then 100 will be given because News is mostly related to Proper nouns than Nouns. If Entity is tagged as NN more time, then 60 will be given and if both counts are equal then 50 will be the weight. These values are assigned on experimental basis because there is no factor of normalization in this particular case.
3) For assigning Occurrence in heading weight firstly count of all Entities in heading is calculated, If there is only one Entity it is given 100, if there are more than one entities then the entity that occurred first in the heading will be given 100, 2nd entity will be given 90 and so on till 50, and in case of more than five entities 50 will be given to rest them because unless they are too much in number that the News cannot discuss these Entities altogether in one News but still some of them have probability of being the important Entity of the News.
4) To give weights to entities w.r.t their occurrence in news, the repetition count of all distinct entities is calculated. Then the top 10 repeated entities are given weights by applying the following strategy:
   - The maximum repeated entity is given a weight of 100 and a Repetition Percentage (RP) of 100%.
   - The other 9 entities are given weight by calculating the percentage of Repetition Count (RC) of proceeding Highest Repeated Entity (HRE) w.r.t to the RC of the preceding HRE. Then this percentage is subtracted from the repetition percentage of the preceding HRE and the half of this value is added to the RP of proceeding highest repeated entity.

Using the following formula in (1):

$$(\boldsymbol{RP})_{\boldsymbol{n}} = \left(\frac{\boldsymbol{100}.(\boldsymbol{RC})_{\boldsymbol{n}}}{(\boldsymbol{RC})_{\boldsymbol{n-1}}}\right) + \frac{\boldsymbol{1}}{\boldsymbol{2}}\left\{\boldsymbol{100} - \left(\frac{\boldsymbol{100}.(\boldsymbol{RC})_{\boldsymbol{n}}}{(\boldsymbol{RC})_{\boldsymbol{n-1}}}\right)\right\} \quad (1)$$

- For example: If the highest occurrence is 20 and second is 15 , then the Repetition Percentage (RP) of second is calculated by taking out the percentage of 15 w.r.t 20 i.e. 75%. Then calculating the difference of 75 and 100 (i.e. the RP of first HRE). The difference comes out to be 25. The half of this is 12.5 and is added to the previously calculated RP of second HRE. So the final RP of second HRE will be 87.5% and weight will be 87.5.
- This factor is added to reduce the difference in weights of two consecutive important Entities.

5) To give weight to entities w.r.t their occurrence/position in first sentence, the first Entity is given 100 and all the proceeding entities are given weights with a difference of 5 descending.
6) To give weight to entities w.r.t their occurrence/position in other sentences of the News corpus the same strategy is applied as for the first sentence except that at the end all weights for each Entity is summed up and divided by one less than the total number of sentences (already assigned weight for first sentence).
7) As it is discussed earlier when there is no match for any word in a lexicon, according to Brill tagger that word is tagged as Noun (NN) and that NN is also considered which will create problem in selection of relevant entities and as a result some irrelevant entities also show up in the result. To overcome this problem Ambiguity Correction weight of 100 is given to all PN and those NN that were present in lexicon and 75 is given to NN that does not exist in lexicon. These figures are eventual result of experiments and research because there is probability of both cases either that word is tagged NN wrongly or there is something new that is NN and not present in lexicon.

A corpus can have multiple Entities but it was analyzed that in one News only few of the Entities are important and their number is not more than three on which the whole News depends. So for that a threshold is defined, if the difference of total weights of two Entities is greater than that threshold then all the Entities having less total weight will be neglected, and not more than three Entities will be considered as important. Only those Entities that qualify in this step will be used in the next step for association with Saliences.

C) Salience Finding

This Section will discuss rule based approach to relate Saliences with Entities that are extracted in the previous section. Salience is those words which decide the polarity of the news or specifically the sentence discussing any Entity. Only those Entities will be used that were declared important in the previous Section. This paper is concerned to know the polarity of News with respect to one or more Entities in the News, so it is desirable to analyze that which Salience is contextually and linguistically connected to which Entity. In other words we can say that the goal is to know Saliences of each particular Entity in the News. It has been analyzed that Salience can only be Adjective, Adverb, Verb or Noun. Although, there are some challenges in this process. Firstly, it is quite difficult to make the computer understand what the context of the News is. Secondly, the structures of sentences in Urdu language are not fixed. However, this paper will describe two interlinked approaches (not independent to each other) to achieve this goal up to great extent.

1) N-Gram:

To implement this the first step is to remove all the words consisting only two characters like: **کو** , **ان** , **کی** etc. These words do not contribute in what we are going to achieve, instead these words wastes a token between, in so many grams, if we do not remove them then some bi-grams will come in tri-grams and some five-grams will be missed, so it will decrease the overall efficiency. We made bi-grams, tri-grams, four-grams and five-grams out of our POS Tagged News corpus for each Entity such that there should be an Entity and a Salience in each N-Gram. Issues and development of N-Grams in Urdu are discussed by Farah Adeeba et al., 2014 [22]. The lesser the value of N the greater is the probability that the word is said for that particular Entity. So to give normalized weights we divide 100 into four parts to assign weights to each N-Gram as 40 for Bi-Grams, 30 for Tri-Grams, 20 for Four-Grams and 10 for Five-Grams. It would hardly a case that any relevant Salience will occur for any Entity in Six-Gram after removing two character words. After this weight assigning, lists of Saliences for each Entity is obtained and then total N-Gram weight is calculated by searching each Entity-Salience pair in every N-Gram and taking the summation of their weights.

2) Linguistic Analysis:

This study discusses Linguistics in a sense that how context is being made in Urdu Language. This technique is applied on sentence basis. The first step is to find the weight, so the normalized technique is to divide 100 by number of Saliences in a sentence. In attempt to find Saliences for each Entity several cases has been considered:

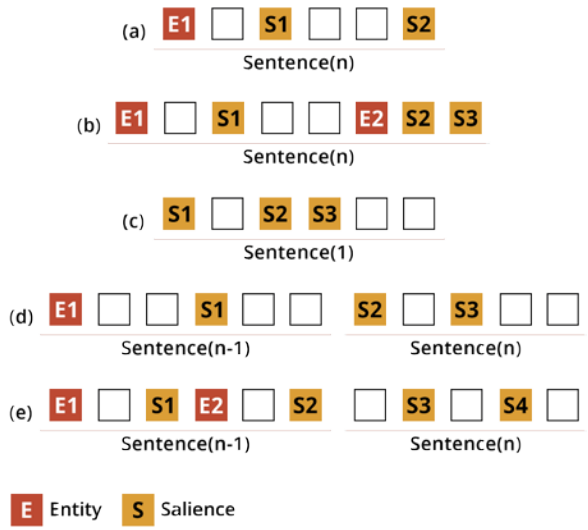

Fig. 2. Salience in sentences

a) If there is only one Entity in a sentence then all the Saliences of that sentence will belong to that particular Entity and sentence wise weight will be assigned to those Saliences. Fig. 2.(a) illustrates this case.
b) As in Fig. 2(b), if there are more than one Entities in a sentence than it will be quite complex to judge that which Saliences belongs to which Entity. For this case N-Grams will be considered as mentioned in the previous technique to assign percentage of the calculated sentence wise weight of the Salience in this technique.
c) If there are only Saliences and no Entity in a sentence as shown in Fig. 2(d) and 2(e), then the Entities of the previous sentence will be considered because whenever we have some words that are defining some polarity and there is no entity to be associated with, so there is a very high probability that a context of previous sentence has been used in the current sentence, or in other words the Saliences are referring the previously discussed Entities. If there is no Entity in the previous sentence then it will go further back to find some Entity, and finally assign sentence wise weight to each Entity-Salience pair. But what if the sentence is the first sentence of the News as shown in Fig. 2(c), then it is considered that the Saliences of that first sentence will belong to those entities which are present in the heading, keeping in mind that it is not necessary that every important Entity should be in heading of the News it might be declared an important Entity because of other factors discussed in previous section, because here is the probability that the context has been taken from heading of the News so we assign sentence wise weight to their Entity-Salience pairs.

At the end of this process overall result of both techniques decides the total weights of Entity-Salience pairs that will be used for assigning polarities. Fig. 3. shows association of different saliences with entities.

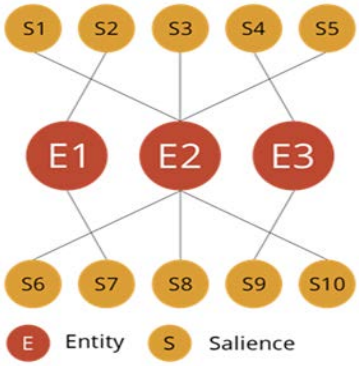

Fig. 3. Entity-Salience Pairs

D) POLARITY ASSIGNING

It is the process of associating polarities from the scale of -10, -5, 0, +5, and +10 (-10 being most negative to +10 being most positive) to each Salience in the corpus. This scale is chosen to mitigate the error in giving a particular Polarity to a word. Software named Manual Polarity Tagger (MPT) was developed to tag Polarity words. Hundred corpuses were tagged; the number of manually tagged word is given in Table 3 Section 4. The existing lists of Polarity words are also used for processing as given in Table 4. of Section 4. The Existing lists were not tagged by MPT for the sake of normalization of results. The words of positive list word were taken as +5 and words of negative list were taken as -5.

In this research polarities are assigned on the basis of linguistic and structural analysis and not on the perspective of any Entity within or outside the given News corpus.

For Example: “X country’s Army fires on Y country” It is Positive for X country but Negative for Y and “X county’s Army fires on X country’s civilians” is Negative for X country but might be Positive for any country Y. So these Perspectives will not be considered and just the context will be noted that firing is negative whoever did that.

E) OVERALL POLARITY CALCULATION

After assigning the polarities the corpus was searched for Intensifiers, Negations and Double Negations [8]. In case of Intensifiers like, شدید , زیادہ or بہت etc. the Polarity of the Salience with an intensifier will be doubled like if the Salience has -5 polarity it will become -10 with that intensifier.

In case of negations like نہ or نہیں etc. the polarity of the Salience will be inverted like if the Salience has -5 polarity it will become +5 with that negation. In case of Double Negation there will be no effect on the polarity of the Salience. After this step Overall Polarity for each

important Entity is calculated in two ways firstly the number of positive and negative Saliences (associated in Section VI) is counted and in the second step total weights of each positive and negative Salience are summed up.
Finally the number of Saliences and total Polarities of each Entity are used to determine the result in Natural Language. If the number of negative Saliences is less than the number of positive Saliences but the total polarity weight of negative Saliences is greater than the total polarity weight of positive Saliences then the Polarity of News for that particular Entity is negative. If the number of negative Saliences and total polarity weight of negative Saliences are greater than number of positive Saliences and total polarity weight of positive Saliences then the Polarity of News for that particular Entity is highly negative. Same is the case for positive Polarity of News for any Entity. If the number of Saliences and total weights of both positive and negative Saliences are equal then it will not simply become neutral but it will tell that the Entity has both polarities. This Software product can also give the track of the News that in which part of the News the polarity is either positive or negative.
For the sake of Verification of our results that will be generated by the software, some corpus were tagged manually for finding Polarity of words using Manual Polarity Tagger developed as part of this research, and then all calculation were made following the steps in this research to find manually the final results that were expected to be generated by the software product. At the end of the research the software generated results were verified by comparing with manual results and also by Analyzing more software generated results manually. After all this process the Software results were found Promising.

IV. Experimental Results

This section will illustrate the working of major processing modules of this research. Fig. 4. is the heading of Corpus we have taken from Nawai Waqt [25] for giving example:

مسلم لیگ کا اخلاقیات اور اقدار سے دور کا تعلق بھی نہیں، محمودالرشید

Fig.4.

Corpus of News detail is given in Fig. 5:

لاہور:پنجاب اسمبلی میں اپوزیشن لیڈر میاں محمودالرشید نے کہا ہے کہ پنجاب اسمبلی سمیت ہر فورم پر حکمرانوں کی کرپشن اور لوٹ مار کو بےنقاب کر یں گے مسلم لیگ کا اخلاقیات اور اصولوں سے دور کا بھی تعلق نہیں انکو سیاسی تر بیت کی ضرورت ہے,غر یبوں کے مسائل کا حل صرف تحر یک انصاف کے پاس ہے جو ملک کو کر پشن,لوڈشیڈ نگ اور دیگر مسائل سے مکمل نجات دلائے گی حکومت مزید مہنگائی کرنے کی بجائے عوام کو ریلیف دینے کیلئے اقدامات کریں۔

Fig.5.

After POS Tagging words that are important for further processing as tagged by POS Tagger are shown in Table 6.

Table 6: Useful Words and their Tags

POS Tag	Words
NN	بےنقاب , کرپشن , حکمرانوں , لیڈر , اپوزیشن , غریبوں , تربیت , تعلق , اصولوں , اخلاقیات , لوڈشیڈنگ , ملک , تحریک انصاف , حل , مسائل , مہنگائی , اقدامات , ریلیف , عوام , حکومت , نجات
PN	مسلم لیگ , محمودالرشید , پنجاب اسمبلی
ADJ	مکمل , سیاسی
VB	دینے , کرنے , دلائے , کریں , لوٹ مار , کہا
ADV	مزید , بجائے , صرف
NEG	نہیں

The POS Tagged corpus then fed to Entity Finding module, Table 7 shows the top 3 entities after applying criteria given in Section 6(B). As discussed in Section 6(B) a threshold is set for considering important entities for further processing. After intensive analysis 15 is set as the threshold value. If the difference of total weights of two consecutive entities is greater than or equal to 15 then the entities of lower total weights will not be considered. In this example 1st and 2nd Entities will be considered for further processing because the difference of total weights of 2nd and 3rd Entities are greater than the threshold value. Further top Entity will be decided on the basis of number Saliences attached with each Entity.

Table 7: Result of Entity Finding module

Word	POS	Heading	Occurrence	1st Sentence	Other Sentences	Ambiguity Correction	Total Weight
مسلم لیگ	100	100	75	65	0	100	104
محمودالرشید	100	90	75	80	0	100	102
پنجاب اسمبلی	100	0	100	90	0	100	84

The Accuracy of this module for assigning weights as compared to manual assignment of weights is 96%. The Accuracy of Finding Correct Entity for a given corpus is 89%. This accuracy value is calculated by processing 100 corpuses manually. Fig. 6. shows manual and software calculation of total weights for this module. Overlapping of graphs shows matching of total weights for each entity.

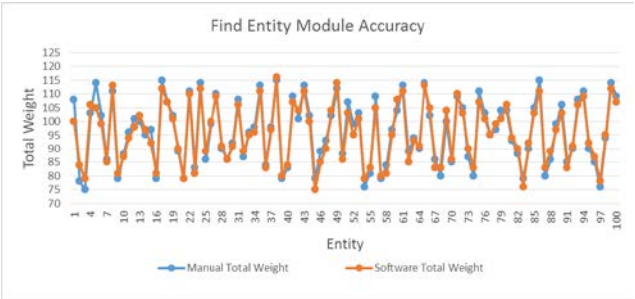

Fig. 6. Entity Module Accuracy

As demonstrated in Fig 6, each Salience will be associated with entities in the corpus with weight assigned on basis of

N-Gram and Linguistic structure. Table 8 shows the Saliences associated with each entity in N-Gram Analysis and Table 9 shows Saliences associated with each entity in Linguistic Analysis.

Table 8: Entity-Salience Pairs by N-Gram

Entity	Saliences
مسلم لیگ	, بےنقاب , لوٹ مار , کرپشن تعلق , اصولوں , اخلاقیات
محمودالرشید	اپوزیشن

Table 9: Entity-Salience Pairs by Linguistic Analysis

Entity	Saliences
مسلم لیگ	, اخلاقیات , بےنقاب , لوٹ مار , کرپشن , تربیت , سیاسی , تعلق , اصولوں , لوڈشیڈنگ , حل , مسائل , غریبوں اقدامات , ریلیف , مہنگائی , نجات , مکمل
محمودالرشید	, بےنقاب , لوٹ مار , کرپشن , اپوزیشن اصولوں , اخلاقیات

In Table 9 after entity **محمودالرشید** another entity **مسلم لیگ** is present so some entities are associated with **محمودالرشید** because of N-Gram but all entities after **مسلم لیگ** are associated with **مسلم لیگ**, this is due to the change of context within a sentence. After assigning individual weights to each Entity-Salience pair, total weight of each pair is calculated considering each individual weight given in both N-Gram and Linguistic Analysis. Table 10 shows final Entity-Salience pairs that will be used for polarity assigning module.

Table 10: Final Entity-Salience Pairs

Entity	Saliences
مسلم لیگ	, اخلاقیات , بےنقاب , لوٹ مار , کرپشن , غریبوں , تربیت , سیاسی , تعلق , اصولوں نجات , مکمل , لوڈشیڈنگ , حل , مسائل
محمودالرشید	اپوزیشن

After analyzing the number of Saliences associated with each Entity further processing will be done on Entity **مسلم لیگ** only. The accuracy of Salience finding module is found to be 84% neglecting errors due to incorrect POS Tagging by POS Tagger.

These pairs and their respective weights are then fed to Polarity Assigning module. Initial weights of these Saliences as tagged by this module are given in Table 11.

Table 11: Initial Polarities of Saliences

Polarity Weight	Saliences
-10	لوٹ مار , کرپشن
-5	, مسائل , غریبوں , سیاسی , بےنقاب اپوزیشن , لوڈشیڈنگ
+5	, حل , تربیت , تعلق , اصولوں , اخلاقیات نجات , مکمل

After this step the whole corpus is processed for Intensifiers, Negations and Double Negations. In this example corpus only case of negation is present. Here we have more than one Salience near and before the negation so, if all those Salience are either positive or negative then polarities of all of them will be inverted, otherwise the polarity of the Salience just before negation is checked and moving backwards invert the polarities of Saliences having same polarity as the first one until a Salience of opposite polarity is reached among those Saliences. Table 12 shows finalized Polarities of Saliences after this step. The accuracy of Polarity assigning module in terms of tagging correct polarity is 80% with the lack of resources of Polarity Tagged words.

Table 12: Final Polarities of Saliences

Polarity Weight	Saliences
-10	لوٹ مار , کرپشن
-5	, مسائل , غریبوں , سیاسی , بےنقاب تعلق , اصولوں , اخلاقیات , اپوزیشن , لوڈشیڈنگ
+5	نجات , مکمل , حل , تربیت

In the last step number of saliences and total weight of polarities will be calculated as shown in Table 13.

Table 13: Number of Salience and Total Weight of Polarity

Polarity	Number of Salience	Total Weight
Positive	4	+20
Negative	11	-65

As shown in Table 13, the number and total weight of negative Salience are greater than the number and total weight of positive Salience. Hence, as discussed in Section 6(E) the overall polarity of the News for Entity **مسلم لیگ** is Highly Negative with an overall accuracy of the system of 84.5%.

V. CONCLUSION

This research presented heuristic approach to finding Saliences in Urdu News corpus by finding Polarity (degree of positivity or negativity) of News for any particular Entity.Online Urdu newspapers are used for processing different steps involved in salience finding are: Parts Of Speech tagging, Finding important entity, saliences in the corpus, assignment and calculation of Polarities for particular Entity using heuristic approach. Based on the total weight of polarity (negative and positive), the accuracy achieved by the system is 84.5%. The results of this research are much promising and accurate up to great extent. But, there are still many stones left unturned in this domain. For future research work, authors are focusing on features extraction for Entity and Salience finding with

more accurate assignment of polarity using more polarity tagged words and performing the Salience analysis using learning machine and computational intelligence based approaches.

References

[1] B. Pang and L.Lee. "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, Vol. 2, pp. 1–135, 2008.

[2] S. Mukund, D. Ghosh and R. K. Srihari. "Using Sequence Kernels to identify Opinion Entities in Urdu", In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 58-67, 2011.

[3] Z. S. Afraz, A. Muhammad and A. M. Martinez-Enriquez. "Sentiment-Annotated Lexicon Construction for an Urdu Text based Sentiment Analyzer" , Advances in Artificial Intelligence - 9th Mexican International Conference on Artificial Intelligence, MICAI, Pachuca, Mexico, Proceedings, Part I , vol. 63 (4), pp. 218 – 221,2010.

[4] Z. S. Afraz, M. Aslam and A. M. Martinez-Enriquez. "Lexicon Based Sentiment Analysis of Urdu Text using SentiUnits", 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, Proceedings, Part 1, Vol. 6437, pp. 32-43,2010

[5] A. Buche, M. B. Chandak and A. Zadgaonkar. "Opinion Mining and Analysis: A Survey", Vol. 2, No.3, 2013.

[6] V. Lakshmi and M.S.Vijaya. "Opinion Mining Using Sequence Labeling", Transactions on Machine Learning and Artificial Intelligence, Val 3.No. 3, pp. 48-50,2015.

[7] W. Anwar, X. Wang and X. L. Wang. "A Survey of Automatic Urdu Language Processing", Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 4489 - 4494, 2006.

[8] J. Bhaskar, K. Sruthi and P. Nedungadi. "Enhanced Sentiment Analysis of Informal Textual Communication in Social Media By Considering Objective Words And Intensifiers", IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), pp. 1 - 6, 2014.

[9] S. Hussain. "Resources for Urdu Language Processing", In the Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08, IIIT Hyderabad, India, 2008.

[10] D. Becker and K. Riaz. "A Study in Urdu Corpus Construction", COLING '02 Proceedings of the 3rd workshop on Asian language resources and international standardization, vol.12, 2002.

[11] A. Muaz, and A. Hussain. "Analysis and Development of Urdu POS Tagged Corpus" ,In the Proceedings of the 7th Workshop on Asian Language Resources, IJCNLP'09, Suntec City, Singapore, pp. 24–31,2009.

[12] Z. Rehman, W. Anwar and U. I. Bajwa. "Challenges in Urdu Text Tokenization and Sentence Boundary Disambiguation", ACL Anthology Network, pp. pages 40–45,2011.

[13] N. Durrani and S. Hussain. – Urdu Word Segmentation – HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, US, 2010.

[14] G. S. Lehal. "A Two Stage Word Segmentation System for Handling Space Insertion Problem in Urdu Script", Proceedings of World Academy of Science, Engineering and Technology, Bangkok, Thailand, Vol. 60, 2009.

[15] G. S. Lehal. "A Word Segmentation System for Handling Space Omission Problem in Urdu Script", Workshop on South and Southeast Asian Natural Language Processing, pp. 43–50, 2010.

[16] T. Brants. "A statistical part-of-speech tagger" , In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000 Seattle, WA, USA, pp. 224-231,2000.

[17] W. Anwar, X. Wang, L. Li and X.L. Wang. "A Statistical Based Part Of Speech Tagger for Urdu Language", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 3418 – 3424, 2007.

[18] B. Megyesi. Brill's Rule Based POS Tagger, 1992.

[19] H. Sajjad and H. Schmid. "Tagging Urdu Text with Parts of Speech: Tagger Comparison" , EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 692–700,2009.

[20] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen, F. Adeeba, and M. Butt. "The CLE Urdu POS Tagset", Poster presentation in Language Resources and Evaluation Conference (LERC 14), Reykjavik, Iceland, 2014.

[21] A. Hardey. "Developing a tagset for automated part-of-speech tagging in Urdu", Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16. Department of Linguistics, Lancaster University, 2003.

[22] F. Adeeba, Q. Akram, H. Khalid, and S. Hussain. "CLE Urdu Books N-Grams", Poster Presentation in conference on Language and Technology 2014 (CLT 14), Karachi, Pakistan, 2014.

[23] U. Singh, V. Goyal and G. S. Lehal. "Named Entity Recognition System for Urdu", 24th International Conference on Computational Linguistics, pp. 2507–2518,, 2012.

[24] M. Elarnaoty, S. AbdelRahman, and A. Fahmy. "A Machine Learning Approach for Opinion Holder Extraction in Arabic Language", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No. 2, 2012.

[25] http://www.nawaiwaqt.com.pk/E-Paper/Lahore/2015-04-09/page-9/detail-0

[26] http://chaoticity.com/urdu-sentiment-lexicon/

[27] http://www.cle.org.pk/Publication/theses/2007/part_of_speech_tagger.pdf.

S. Abbas Ali, Ph.D. in Computer Science (Automatic Speech Recognition & Machine Leraning). Research interest includes: computational intelligence and pattern recognition.

Muhammad Daniyal Noor, B.E from NED University of Engineering and Technology. Research interest includes AI, NLP and Software Engineering.

Munir Ahmed Javed, B.E from NED University of Engineering and Technology. Research interest includes AI and Human Computer Interaction.

Muhammad Mohsin Aslam, B.E from NED University of Engineering and Technology. Research interest includes ML and NLP**.**

Omer Ahmed Khan, B.E from NED University of Engineering and Technology. Research interest includes ML and Big Data.

Noor Us Sahar Zubair, B.E and M.E from NED University of Engineering and Technology. Research interest includes ML, Big Data Analytics, Networking and Cloud Security.