Diksha Goyal, Gurpreet Singh Josan

Automatic Sentiment Lexicon Construction For Punjabi

[1]Diksha Goyal, [2]Gurpreet Singh Josan

[1]Department of Computer Science Punjabi University, Patiala
Email Id: dkshgyl@gmail.com

[2]Department of Computer Science Punjabi University, Patiala
Email Id: josangurpreet@pbi.ac.in

ABSTRACT—Sentiment Analysis has become a revenue generation model. The backbone of any Sentiment Analysis is Sentiment Lexicon. Using the available sentiment lexicon to develop new sentiment lexicon in other language is an interesting area of study and focus of this paper. Available resources like Hindi SentiWordNet, English SentiWordNet and Punjabi WordNet are used to prepare the sentiment lexicon for Punjabi Language with their positive and negative scores and IDs. The prepared dataset includes unique entries to improve the reliability of Lexicon. This prepared dataset recursively collects synonyms to expand the sentiment lexicon. In the experimental result, it is proved that developedPunjabi Sentiment Lexicon helps in improving the sentiment analysis task or opinion mining task.
Keywords: SentiWordNet, Punjabi Sentiment Lexicon, Punjabi WordNet, Sentiment Analysis, English SentiWordNet, Hindi SentiWordNet, Dictionary.

I. INTRODUCTION

Sentiment Analysis has high demand in market. Famous online shopping websites like Amazon, Naaptol etc. uses Sentiment Analysis to generate revenues. Sentiment Analysis has become a revenue generation model. The backbone of any Sentiment Analysis system is Sentiment Lexicon. Sentiment Lexicon for well-known languages like English, Hindi, [1]Arabic, Punjabi, Bengali and Tamil are freely available to use. Due to advancement in technology, customer reviews are available in regional language like Hindi, Tamil, Bengali, Punjabi, Arabic, and English. To analyze these reviews, such sentiment analysis are required for regional languages also. Researchers developed Sentiment analysis for the above define languages. For such Sentiment Analyzer, Lexicon is the foremost requirement. Availability of good lexicon will definitely improve the result of lexicon system. Development of good lexicon needs either lots of human efforts or requirement of annotated corpus to develop machine learning and statistical based system. Use of Sentiment Lexicon, to develop Sentiment Lexicon in other language is an interesting area of study and the focus of this work. This paper explains the process of developing Punjabi Sentiment Lexicon derived from Hindi SentiWordNet and English SentiWordNet. The approach to know about the sentiment in text present in literature, concerns the use of dictionary resources like SentiWordNet and synset list.

In [2], several computational techniques to generate Sentiment Lexicons in Punjabi Language automatically have been discussed. Several prior polarity sentiment lexicons are available for English such as [3]English SWN (SentiWordNet), for Hindi such as [4]HSWN, [5]Hindi SWN. This paper makes the use of English SWN and HSWN. Due to lack of weightage and continuous scoring in Hindi SWN, it is not being used.

Among these freely available Sentiment Lexicon resources we find that SentiWordNet is widely used in several applications like Sentiment Analysis, opinion mining and emotion analysis. SentiWordNet extricates data from WordNet database and change it into an opinion mining information which is made freely available[6].

Over the previous years it was difficult to summarize the huge amount of data into a single form. To solve this problem, many resources are now available like SentiWordNet, Lexicon which are modeled for human use to summarize the annotated data. By getting inspired from this, it is a small step to develop a Sentiment lexicon for Punjabi language which describes the positive and negative polarity of content. It will help to summarize the large data into positive and negative scores.

Punjabi is the language used by hundreds of millions of people in India and Pakistan. Punjabi is an Indo-Aryan language[7]. It is the native language of about 130 million people, and is the 10th most spoken language in the world. Surprisingly, a few work has been done in lexical field in Punjabi language. Motivated from this, it was decided to develop a sentiment lexicon for Punjabi language.**WordNet** is a lexical database which defines the set of words with their synonyms and other sort of information about the word. Mainly it is used to express the similarity between the words. It is basically a knowledgebase where information, properties and linkages are stored. WordNet is basically a database in which set of words are grouped with their one or more synonyms without missing the true value of prepositions in which they are embedded. This set of synonyms is called **Synset. SentiWordNet** is basically the combination of sentiment information and WordNet. For each WordNet synset s, SentiWordNet assigns three sentiment scores: Positive (Pos), Negative (Neg), Objective (Obj).

For example, the scores for the word ਅਗੋਚਰ (agōcar) are:

$$\text{Pos (ਅਗੋਚਰ)} = 0.125$$

$$\text{Neg (ਅਗੋਚਰ)} = 0.625$$

Objective score can be calculated by[8]

$$\text{Pos (s)} + \text{Neg(s)} + \text{Obj(s)} = 1$$

ROLE OF LEXICON IN NATURAL LANGUAGE PROCESSING

Lexicon is basically a part of NLP, which plays a major role in each task of NLP. It contains the data about words or word strings. For almost all of the NLP tasks, lexicon is like backbone. For example lexicon is required in Machine Translation, Machine Transliteration, Named Entity Recognition, Information Retrieval, Word sense Disambiguation, Text Summarization, Speech Recognition and Speech Synthesis etc.

By improving the effectiveness of Supervised Learning Approach to Sentiment Analysis, Lexicon improves the result of Sentiment Classification along with performing an important task in Sentiment Analysis. It is the basic requirement for vocabulary.

NEED OF PUNJABI SENTIMENT LEXICON

Based on literature survey up to date, in the Sentiment analysis there is not much work on the Punjabi Sentiment Analysis and on the lexicon construction for Punjabi Language. We aim to develop Sentiment Lexicon for Punjabi Language. There are sentiment lexicons available for other languages but there is no such lexicon available for Punjabi language. Therefore we proposed a lexicon with polarity scores for Punjabi language by using English SWN, Hindi SWN, Hindi-Punjabi Dictionary, English Punjabi Dictionary and Punjabi Synset from Punjabi WordNet. Online Punjabi novel, Punjabi literature etc. are the main sources from where we collected the dataset.

It was very difficult to easily access the opinion of customers. Customer usually makes purchase by reading the reviews. As posts increasing day by day that makes it troublesome for a client to condense it. To handle this type of trouble, Punjabi sentiment lexicon has been constructed, a word reference technique which is used nowadays and will generally be used in future.According to [9], Sentiment Lexicon generation is divided into various approaches: Manually based approach, Dictionary based approach and Corpus based approach. Manually based approach is one in which human works manually to develop any resource. Manual development is moderate, costly, inconvenient, unmanageable and tedious to build and upgrade by hand. Another one is dictionary based approach, in which a seed list has been expanded with the help of existing resources like WordNet, SentiWordNet etc. This automatic construction is fast, less expensive, high potential and high performance. Third one is corpus based approach, in which manually labelled seed words are used based on the corpus data. Among all the above approaches dictionary based approach has been followed.

The rest of this paper is organized as follows: Section 2 presents the survey of the paper. Section 3 describes the approach used with algorithms and flowchart in detail. Section 4 shows the Experimental Analysis with description of collected data, Selected Features. Section 5 shows the evaluation and result. Lastly we conclude this paper and discuss the future work in Section 5.

II. ANALYSIS OF EXISTING WORK

[10] by making the utilization of WordNet, Author represents the method for build lexicon of input words. Author tested framework through various domains like science, agriculture, sports etc. each containing about 800 words. Results meet the accuracy for nouns is 93.9%, for verbs 84.4%, for adjectives 72.4% and for adverbs 58.1%. Techniques used are Word Sense Disambiguation, an inference, and the knowledge base of the Universal Networking Language (UNL). [11] developed a method for the automatic construction of an Arabic Lexicon. Accuracy is achieved to about 96%. To receive any data in Arabic language, creator defines many rules that extract the linguistic attributes. Several utility processes are implemented like a pattern extractor, a stemming process and a part of speech tagging process. [12] describes a SentiWordNet(version 1.0) which is an opinion lexicon derived from WordNet database where each term is associated with numerical scores, indicating objective, positive and negative terms are contained in the synset. Techniques used are Rocchio and Support Vector Machine.[13]proposed an approach can make full use of the emphasizing between documents and words. Three domain specific datasets are taken: Hotel reviews, Stock reviews and

electronic product reviews. Approaches used are Thesaurus based and raw corpus based. [14] Discuss the improvement of SentiWordNet 3.0 over SentiWordNet 1.0. It reports the result of calculating SentiWordNet 3.0 over WordNet 3.0. SentiWordNet 3.0 is more accurate from SentiWordNet 1.0 with 19.48% improvement for positive and 21.96% improvement for negative. The result was founded as 20% improvement w.r.t. SentiWordNet 1.0. [15] proposed to build a sentiment lexicon which is domain independent. They present a Machine Learning Based Sentiment Lexicon by extracting data from Amazon corpus from different domains. This gain basically improvement over Automatic Build Lexicons like SentiWordNet. Input dataset has been taken from Amazon Product reviews data set, Movies data. Techniques used are Support Vector Machine. [16]proposed a novel probabilistic modelling framework which is known as Tag Sentiment Topic Model (TSTM) to construct a lexicon for sentiment analysis task. It is based on Latent Dirichlet Allocation (LDA). Input data is Movie Review Dataset consisting positive and negative review. It covers the work in two parts: Sentiment classification and lexicon construction.[17]purposed to build a Subjective lexicon for Hindi. Author presented the development of building up a dictionary of adjectives and adverb utilizing Hindi WordNet and building up corpora of Hindi Product Reviews. This paper achieved 70.4% agreement with human annotators and 79% approx. correctness on product review classification. For item review dataset in Hindi language they deciphered pre-annotated data on Amazon item surveys from English to Hindi using Google. All the interpreted surveys were of length <=25. Technique used are WordNet and Breadth first search traversal.[18] has main focus to learn a domain specific Sentiment Lexicon. Experiment has been done on hotel review dataset and customer feedback surveys on printers identify different polarities of a word depend on the aspect in context. [19] developed the Subjective Lexicon for Indian languages by using the Hindi Subjective Lexicon. An algorithm is designed by combining the simple scoring method and the unigram method. Method used is Bi-Lingual dictionary helps to provide the translation process at word level. [20] used WordNet to construct a subjective Lexicon. Author gives a subjective lexicon based on Sinhala language (spoken by people in Sri Lanka). Dictionary for Sinhala Language has been produced with the guide of English estimation lexicon by using English SentiWordNet 3.0. Result is acceptable maximum of 60% in Naïve Bayes Classification. Techniques used are Naïve Bayes, Support Vector Machine. [21] built a thesaurus Lexicon for Sentiment classification. This is basically a Dictionary based approach and makes the use of three online dictionaries.

III. PROPOSED APPROACH

Algorithm proposed in this research is dependent on the pre-annotated [22]Punjabi WordNet and SentiWordNet. It is assumed that synonyms have the same polarity in HSWN and Punjabi Synset as their root words have. *Table1* shows the information about collected pre-annotated data. Punjabi-Hindi dictionary and English-Punjabi Dictionary available at Department Of Computer Science, Punjabi University, Patiala

Using the method explained here, we constructed a sentiment lexicon for Punjabi language.

1.) Initially, we collected 5048 total entries which were common in both Punjabi-Hindi dictionary and HSWN. The Punjabi translation of these Hindi word entries was done via Hindi-Punjabi Dictionary and the polarity of that entries was taken from HSWN.
2.) Similarly, we made the use of English-Punjabi Dictionary and ESWN which extended the list by adding 8,163 more entries.

3.) To further extend the list, Punjabi WordNet is used which provides 5100 more entries. If root word exists in the final list but synonyms do not, then give same score to synonyms and add it to the final list. If synonyms exist in the list but root word does not, then add root word with same score as synonym.
4.) To further extend the list, rootword list is used which includes multiple inflected words derived from root words. The root words of an inflected word by given the same score as an infected. This provides 700 more entries to the list.

This is how we got a list of 19,011 unique entries with the help of Punjabi WordNet, HSWN, Punjabi-Hindi Dictionary, ESWN, and English-Punjabi Dictionary. Every entry of sentiment lexicon is associated with positive and negative score. *Table 2* shows the structure of proposed lexicon.The defined terms are associated with two numerical scores, each indicating the positive and negative bias. For more detail about collected data refer to *Table 3*.

OPINION REVIEW DATASET

Punjabi Tribune is famous Punjabi newspaper. Readers Opinion related to current scenario has been published in these newspaper under column '*Pathkan De Khat*' (Reader'scolumn) from Punjabi Tribune and 5abi.com.

Human Annotation: Initially, Total 86 documents were collected from well-known Newspaper 'Punjabi Tribune'. All of the documents were given to 5 human judges. All the Human Evaluators are experts in Punjabi language and they were asked to label the documents by positive and negative tags. Results were 43 documents for positive and 43 documents for negative.

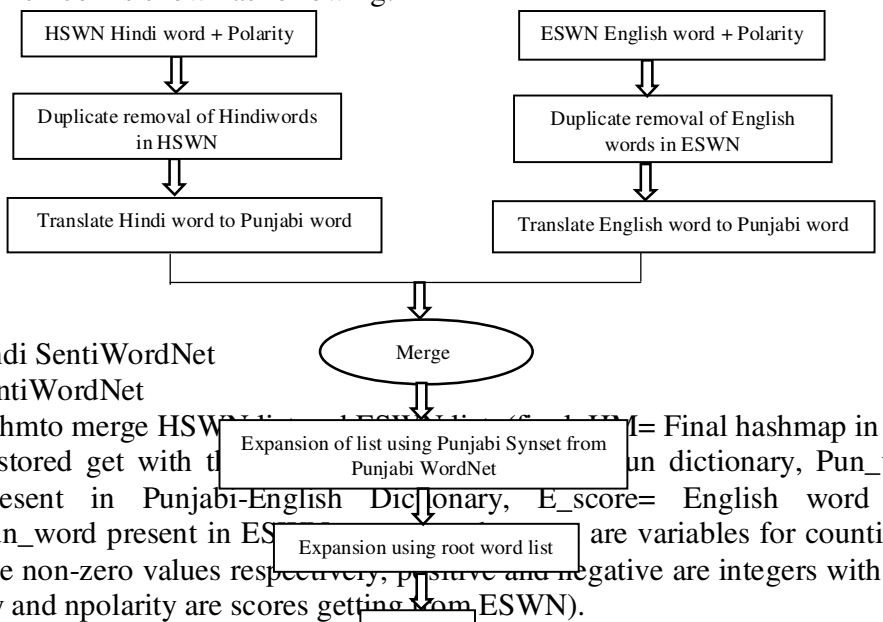
Amazon is a well-known product shopping site. Amazon product reviews collected from 'amazon.in'. As they all were in English language, with the help of Google Translate data was translated from English to Punjabi.

H*uman Annotation:* Initially, Total 80 documents were collected from well-known Amazon. All of the documents were given to 5 human judges and they were asked to be label the documents by positive and negative tags. Results were 40 documents for positive and 40 documents for negative. For more detail refers to *Table4.*

Initially, the Hindi words, Positive and Negative score has been collected from Hindi SentiWordNet. Redundancy of Hindi words are removedby addingthe polarities of all the redundant entriesand inserted into final hash map.This proposed lexicon overcome the drawback of HSWN that a single word has multiple IDs and multiple score. For example the word 'विपरीत' with ID 3636,4473,21973.2497 which creates confusion to choose correct one. Punjabi Sentiment Lexiconony counted the non-zero values of positive and negative score. Now we get a list of Hindi words with scores without any replication. Hindi-Punjabi Dictionary were used to translate Hindi words to Punjabi words.

Parallely, English words, Positive and Negative score has been collected from English SentiWordNet. Redundancy of English words are removed by adding the polarities of all the replicated and inserted into final hash map. It only counted the non-zero values of positive and negative score. Now we get a non-redundant list of English words with scores. English-Punjabi

Dictionary were used to translate English words to Punjabi words.A detailed flowchart of proposed Sentiment Lexicon is shown as following:



where, HSWN= Hindi SentiWordNet
ESWN= English SentiWordNet

Algorithm 1 Algorithmto merge HSW= Final hashmap in which Punjabi words are stored get with t un dictionary, Pun_word= Punjabi words present in Punjabi-English Dictionary, E_score= English word score corresponding to Pun_word present in ES are variables for counting the positive and negative non-zero values respectively, positive and negative are integers with initial value zero, ppolarity and npolarity are scores getting from ESWN).

```
if(!final_HM.Containskey(Pun_word))
        put in final_HM (Pun_word, E_score, countp, countn)
        else
        score =get final_HM(Pun_word)
if(score!=0)
positive=positive+ppolarity;
                        negative=negative+npolarity;
                        countp=countp+cntp;
                        countn=countn+cntn;
put in final_HM(P_word, positive, negative, countp, countn, ID)
        end if
end if
```

In the above algorithm,both the hashmaplists accessed fromHSWN and ESWN were merged by putting into final hashmap. Now, list has been further expanded by using Punjabi WordNet.

Algorithm 2 Algorithm is to expand the list by using Punjabi WordNet. (Punj_syn= Punjabi Synset List, P1_word = Punjabi word in Punjabi WordNet, Synonym= words with similar meaning corresponding P1_word, *Pinput* is an input text document of Punj_syn).

```
P1_word =get Punj_syn(Pinput.txt);
Synonym =get Pun_syn(Pinput.txt);
Score= final_HM.get (P1_word);
if (final_HM . Containskey (P1_word))
```

```
        put in final_HM (Synonym, Score, countp, countn, ID);
else
        if(final_HM. Containskey (Synonym))
        pol= final_HM.get (Synonym);
          if(pol!=0)
                pol=pol+polarity;
               countp++;
                 countn++;
        //pol=<pos, neg, countp, countn, ID>
        Put in final_HM (P_word, pol, countp, countn, ID);
        end if
end if
```

Firstly if root word of Punjabi WordNet exists in the final list but synonyms do not then same score was given to synonyms and added to the final hash map. If synonyms exist in the list but root word does not, then root word was added to the final list with summation of scores of the synonyms.

Algorithm 3 Algorithmto expand the list by using rootword list.(rootword= root word list, rootw= root words present in the list, iword= inflected word corresponding root words, root_word.txt= rootword is the input text document)

```
rootw=get rootword(root_word.txt)
iword=get rootword(root_word.txt)
if(!final_HM.contains(rootw) && final_HM.contains(iword))
        final_HM.put(rootw,final_HM.get(word))
```

Punjabi rootword list is used which includes multiple inflected words derived from root words. If rootword is not in the list but its inflected word is, then add it to final list by giving the same score as its inflected word.

Example: ਝੂਟਾਂਗੀਆਂ (jhūṭāṅgīāṃ) ਝੂਟ (jhūṭ), where ਝੂਟਾਂਗੀਆਂ is the inflected word and ਝੂਟ is the root word. This list contains total 39797 entries.

Algorithm 4 Algorithmto take the average of scores.

```
final_HM.getvalue()
pos=pos/countp;
neg=neg/count;
final_HM.putvalue(ID, pos, neg)
```

Above algorithm calculate the average of polarity scores by taking third and fourth argument of final list as positive and negative non-zero score respectively frequency of each word.ID is a unique numerical value given to each word, word and its synonyms have the same ID.

The proposed algorithm generates score better than the corresponding entry in Hindi and English SentiWordNet. *Table 5*shows some of the entries from Punjabi Senti Lexicon along with corresponding entries from Hindi and English SentiWordNet. The entry in Punjabi lexicon has better score as compared to their counterparts. The score of entry in Punjabi lexicon is generated by averaging all the occurrences of corresponding words in English and Hindi SentiWordNet. Thus, the generated lexicon provides better scoring as also proved by experimental setups.

IV. EXPERIMENTAL SETUP

DATA COLLECTION

Data is collected from 'Punjabi tribune E-newspaper' and 'Amazon online product shopping website'. From the amazon site, product reviews has been accessed. As they were in English language, Online Google translator has been used to translate whole reviews in Punjabi language. Punjabi Tribune dataset has been collected from the '*Pathkan De Khat*'(Reader's Column).

GOLD STANDARD

Dataset has been collected from two sources: Punjabi Tribune and Amazon. Manual approach is the one which includes the human labelling to evaluate the data and to find the accuracy. Initially 100 documents were collected form the Punjabi Tribune randomly. The collected dataset from Punjabi Tribune has been manually checked and tagged by 5 Human evaluators.We asked evaluators to tag the each word as the scale of Positive and negative. After human tagging, total 86 documents (43 positive, 43 negative) were selected out of 100 document and rest were rejected because of ambiguity and disagreement among evaluators regarding the sentiment of document. Similarly, Initially 100 documents were collected from the Amazon website randomly. The collected dataset from Amazon website has been manually tested by 5 human evaluators. After Human tagging, total 80 reviews (40 reviews for positive and 40 reviews for negative) out of 100 and rest were rejected because ambiguity and disagreement among evaluators regarding the sentiment of document. The *Table 6* shows the collected dataset information.

FEATURE SELECTION

Sentiment Lexicon has been evaluated using Machine Learning Approach. Different type of feature set has been selected. Experiment has been done using below feature sets. In Set-I, bag of words feature set is used with Tf-idf values. In Set-II, bag of words along with positive and negative score feature sets are used.

Set-I: In the first experiment, the base model 'bag of words' has been taken as a feature set indicating text as a bag of its words.Tf-idf values of words are used as values of feature. Higher the frequency of a word in document, higher the importance of that word in that document. Tf (Term Frequency) measures as the frequency of each word in a document (f_w) divided by total number of words in a document (N_n).

$$Tf_{m,n} = \frac{f_{m,n}}{N_n} \quad (1)$$

Idf (Inverse Document Frequency) is computed by dividing the total number of documents in which word (w_m) occurs.

$$idf_m = \frac{\log |P_m|}{|N_n w_m \in N_n|} \quad (2)$$

Tf-idf is the combination of two components, Tf and idf. It calculates the composite weightage of each word present in a document. Higher the occurrence of word in all documents lower the value of tf-idf, lesser the occurrence of word in all documents higher the value of tf-idf will.

$$Tf_{m,n}\text{-}idf_m = Tf_{m,n} \times idf_m \quad (3)$$

Set-II: In the second experiment, bag of words along withpositive score (S_p) and negative score (S_n)of document are used as a feature. Positive score has been calculated as the total of all the positive values corresponding each matching word in a document. S_p is the positive score of a document,n is the total number of matching words in a document, $\mathbf{p}(\boldsymbol{w_i})$ is the positive score of a word.

$$S_p = \sum_{i=0}^{n} p(w_i) \quad (4)$$

Negative score has been calculated as the total of all the negative values corresponding each matching word in a document. S_n is the negative score of a document, n is the total number of matching words in a document, $n(w_i)$ is the negative score of a word.

$$S_n = \sum_{i=0}^{n} n(w_i) \quad (5)$$

Weka tool has been used for this experiment. Basically this experiment is showing how much the proposed Lexicon is accurate as comparative to Naive Bayes Classifier and Logistic Classifier.Naive Bayes classifier is basically a generative classifier whether Logistics is a discriminative classifier.

V. EVALUATION AND RESULTS

An experiment has been performed for testing the usability of the constructed sentiment lexicon for Punjabi language. A domain independent dataset has been collected from an E-Commerce site named Amazon and Punjabi Tribune, a famous newspaper as discussed in above section.

SENTIMENT LEXICON PERFORMANCE

Sentiment Lexicon performance has been evaluated by classifying the collected data from Punjabi Tribune and Amazon. Each word of the document is given a particular score with the help of Sentiment Lexicon. The Positive and Negative score of whole document is collected and the average of positive and negative score is taken separately. The finding result has been shown in *Table 7.*Asaccuracy of other lexicons in literature is also in between 60%-70%. It provesproposed Sentiment Lexicon is quite useful.

MACHINE LEARNING BASED EVALUATION

To further substantiate our claim that our lexicon is quite useful, we also experimented using Machine Learning Technique. The performance of Sentiment Lexicon was also tested using Naive Bayes Classifier and Logistic Classifier. Four parameters are calculated by both Naïve Bayes and Logistic Classifier. With the Punjabi Tribune dataset, training file contains 15,816

tokens without polarity entries and 7,774 tokens with polarity entries.With the Amazon dataset, training file contains 10,677 tokens without polarity entries and 5,567 tokens with polarity entries. Experimental result of Set-I and Set-II as shown in *Table 8*.

It is clearly shown in *Table 8*that Logistic classifier for Punjabi Tribune gives 83.7% accuracy when Sentiment Lexicon is used and 82.5% accuracy when Tf-idf value is used. Similarly, Naive Bayes for Amazon dataset gives 89.1% with Sentiment Lexicon and 81.2% with Tf-idf values. It is proved that Punjabi Sentiment Lexicon performance is good.

VI. CONCLUSION AND FUTURE SCOPE

This paper has main focus to generate a Sentiment Lexicon For Punjabi language by merging HSWN and ESWN and offer averaging of the redundant entries score and reweight each entry. The resulting model can be viewed as PunjabiSentiment Lexicon with the reduction of problems caused from duplicacy. The prepared model includes unique entries to improve the reliability of the Sentiment Lexicon. The experimentation explored that proposed Sentiment Lexicon performs good by using Naive Bayes and Max-Entropy Classifiers.In the future works, the list can be expanded by adding slang or informal words which are used in comment section of online shopping websites, TV dramas etc. Lexicon can be further improved by adding antonyms. Techniques for such improvement might be explored in future.

REFERENCES

[1] N. Omar, M. Albared, A. Q. Al-shabi and T. Al-moslmi, "Ensemble of Classification Algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews," *International Journal of Advancements in Computing Technology(IJACT)* , vol. 5.14, p. 77, 2013.

[2] A. Das and S. Bandyopadhyay, "SentiWordNet for Indian Languages," pp. 56-63, 2010.

[3] A. Esuli and F. Sebastiani, 1 June 2010. [Online]. Available: http://www.cs.unh.edu/~cmo66/class_websites/cs405/assignments/a7/wordnet.txt.

[4] M. Sharan, "hINDIswn," 14 April 2016. [Online]. Available: https://github.com/smadha/SarcasmDetector/blob/master/Hindi%20SentiWordNet/HSWN_WN.txt.

[5] A. Das. [Online]. Available: http://www.amitavadas.com/sentiwordnet.php.

[6] B. Ohana and B. Tierney, "sentiment classification of reviews using sentiwordnet," in *In 9th. it & t conference*, Dublin, Ireland, 2009.

[7] "Punjabis," Wikipedia the free encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Punjabis.

[8] S. Ahire, "A survey of Sentiment Lexicons," 2014.

[9] L. Z. and B. L. , Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012, p. 167.

[10] N. Verma and P. Bhattacharyya, "Automatic Lexicon Generation through WordNet," in *GWC*, Brno, Czech Republic, 2004.

[11] G. k. Riyad Al-Shalabi, "Constructing an automatic lexicon for Arabic language," *International Journal of Computing & Information Sciences,* vol. 2.2, pp. 114-128, 2004.

[12] A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource For Opinion Mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, Genoa-Italy, 2006.

[13] W. Du, S. Tan, X. Cheng, X. Yun and Tan, "Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon," in *Proceedings of the third ACM international conference on Web search and data mining*, New York, USA, 2010.

[14] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Language Resources and Evaluation (LREC)*, Valletta Malta, 2010.

[15] A. Hamouda, M. Rohaim and M. Marei, "Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis," *Journal of advances in information technology,* vol. 2.4, pp. 199-203, November 2011.

[16] R. Xie and C. Li, "Lexicon Construction: A Topic Model Approach," in *2012 International Conference on Systems and Informatics (ICSAI)*, China, 2012.

[17] A. Bakliwal, P. Arora and V. Verma, "Hindi subjective lexicon: A lexical resource for hindi polarity classification," in *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, 2012.

[18] Y. Lu, M. Castellanos, U. Dayal and C. Zhai, "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach," in *20th international conference on World Wide Web*, Hyderabad, 2011.

[19] A. K. and V. G. , "Proposed Algorithm of Sentiment Analysis for Punjabi Text," *Journal of Emerging Technologies in Web Intelligence,* vol. 6.2, pp. 180-183, 2014.

[20] N. Medagoda, S. Shanmuganathan and J. L. Whalley, "Sentiment Lexicon Construction Using SentiWordNet 3.0," in *11th International Conference on Natural Computation (ICNC)*, 2015.

[21] S. Park and Y. Kim, "Building thesaurus lexicon using dictionary-based approach for sentiment classification," in *Software Engineering Research, Management and Applications*

(SERA), 2016.

[22] "IndoWordNet," Center for Indian Language Technology, [Online]. Available: http://www.cfilt.iitb.ac.in/indowordnet/.

Table 1: Different resources and total number of words presentTable 2: Structure of SentimentLexicon

RESOURCES USED	TOTAL WORDS
Punjabi-Hindi dictionary	73,664
HSWN	11,929
English-Punjabi Dictionary	22,614
English SentiWordNet	117,659
Punjabi WordNet	52,257

RESOURCES USED	WORDS EXTRACTED
Punjabi-Hindi Dictionary and HSWN	5,048
English-Punjabi Dictionary and ESWN	8,163
Punjabi WordNet	5,100
Root word list	700
Total	19,011

Table 3: Detailed Information about Sentiment `

Table 4: Punjabi Tribune Dataset

RESOURCES USED	WORDS EXTRACTED
Punjabi-Hindi Dictionary and HSWN	5,048
English-Punjabi Dictionary and ESWN	8,163
Punjabi WordNet	5,100
Root word list	700
Total	19,011

TRAINED FILES	TOTAL TOKENS OF PUNJABI TRIBUNE	TOTAL TOKENS OF AMAZON
Without polarity	15,816	10,676
With Polarity	7,774	5,567

Lexicon Entries

Table 5: Comparative scores of some entries in Punjabi Lexicon and English and Hindi SentiWordNet

ESWN	Positive Score	Negative Score	HSWN	Positive Score	Negative Score	Punjabi Senti Lexicon	Positive Score	Negative Score
Good	1	0	बढ़िया	0.625	0.0	ਵਧੀਆ (vadhīā)	0.6174 2425	0.1875

Somewhere	0	0.625	कहीं	0.25	0.125	ਕਿਤੇ (kitē)	0.25	0.375
Less	0.125	0.375	कम	0.0	0.375	ਘੱਟ (ghaṭṭ)	0.2	0.4166 6666

Table 6: Gold Standard Sentiment Data Set

Table 7: Performance Results using Sentiment Lexicon

	Punjabi Tribune Data set	Amazon Dataset
Total Documents	100	100
Predicted Positive	43	40
Predicted Negative	43	40
No Matching	14	20

	Punjabi Tribune Dataset	Amazon Dataset
Total	86	80
Positively tagged documents	29	37
Negative tagged documents	24	24
Precision	60.41%	69.8%
Recall	67.44%	92.5%
Fscore	63.70%	79.56%
Accuracy	61.62%	76.25%

Table 8: Performance of Naive Bayes and Logistic Classifier

Punjabi Tribune Dataset				
	With Polarity		Without Polarity (Tf-idf)	
Classifiers	Naive Bayes	Logistic	Naive Bayes	Logistic
Precision	74.6%	83.8%	74.6%	82.7%
Recall	74.4%	83.7%	74.4%	82.6%
F-score	74.4%	83.7%	74.4%	82.5%
Accuracy	74.4%	83.7%	74.4%	82.5%
Amazon Dataset				
	With Polarity		Without Polarity (Tf-idf)	
Classifiers	Naive Bayes	Logistic	Naive Bayes	Logistic
Precision	89.6%	92.9%	81.4%	91.3%
Recall	89.1%	91.5%	81.3%	91.3%
F-score	89.2%	91.5%	81.2%	91.2%

Accuracy	89.1%	91.4%	81.2%	91.2%

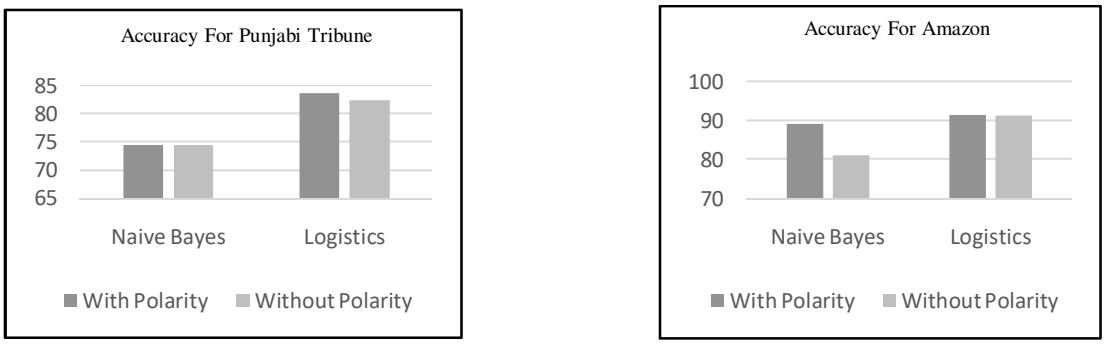


Fig. 1. Chart Representation of Results