# Sentiment Summerization and Analysis of Sindhi Text

Mazhar Ali[1],[2]
[1]Shaheed Zulifqar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi,
[2]Benazir Bhutto Shaheed University, Karachi,
Sindh Pakistan

Asim Imdad Wagan[3]
[3]Mohammad Ali Jinnah University Karachi,
Sindh Pakistan

***Abstract*—Text corpus is important for assessment of language features and variation analysis. Machine learning techniques identify the language terms, features, text structures and sentiment from linguistic corpus. Sindhi language is one of the oldest languages of the world having proper script and complete grammar. Sindhi is remained less resourced language computationally even in this digital era. Viewing this problem of Sindhi language, Sindhi NLP toolkit is developed to solve the Sindhi NLP and computational linguistics problems. Therefore, this research work may be an addition to NLP. This research study has developed an own Sindhi sentimentally structured and analyzed corpus on the basis of accumulated results of Sindhi sentiment analysis tool. Corpus is normalized and analyzed for language features and variation analysis using DTM and TF-IDF techniques. DTM and TF-IDF analysis is performed using n-gram model. The supervised machine learning model is formulated using SVMs and K-NN techniques to perform analysis on Sindhi sentiment analysis corpus dataset. Precision, recall and f-score show better performance of machine learning technique than other techniques. Cross validation techniques is used with 10 folds to validate and evaluate data set randomly for supervised machine learning analysis. Research study opens doors for linguists, data analysts and decision makers to work more for sentiment summarization and visual tracking.**

***Keywords*—Sindhi NLP; sentiment structurization; sentiment analysis; supervised analysis**

## I. Introduction

Supervised classification is important and noteworthy technique of data mining [1], [2] to analyse the text. The supervised classification model works on basis of training and test sets. Training set is used to train the model whereas test set is processed to evaluate the model performance. This research study has developed supervised machine learning model using SVMs. Random Forest and k-NN techniques to identify the true and false classified data from Sindhi structured and sentimental text corpus. Sindhi text corpus is annotated with NLP features therefore, it is multi-class text corpus. The corpus is constructed on basis of accumulated results of Sindhi NLP tool for Sindhi text sentiment analysis. The NLP toolkit is developed for solutions of computational linguistics and NLP problems of Sindhi language. This study verifies the annotation accuracy of Sindhi NLP tool and assesses the performance of machine learning supervised classification model. Supervised model [3] validates the Sindhi text and evaluates it through test dataset. The purpose of this research study is to identify and analyse the Sindhi text sentiment structurization and sentimentally analysed data. Structurization has been done on the basis of five Ws. The questions which are asked in form of five Ws, clear the context of text. No proper research work has been found on Sindhi text structurization for sentiment analysis, therefore, this work generates new path for research on Sindhi corpus for sentiment analysis, semantic analysis, corpus analysis [4] and other linguistics features analysis.

Nowadays, sentiment analysis technique is growing as large number of organizations focus on reviews, sentiments and opinions of people for polarity analysis [5]. Sentiment analysis method is one of the significant methods of NLP [6]. Sindhi text is morphological rich and grammatically complex [7] and users of Sindhi language are settled all over the world [8] thus to work on Sindhi text corpus for sentiment analysis and structurization enable Sindhi users to express their reviews and opinions as well as provide organizations with information to evaluate the sentiments and opinions. Sentiment structurization [9] clarify the status and history of sentiments and helps in tracking the sentiments summaries.

## II. Sindhi Text Structurization for Sentiment Analysis

Sentiment analysis performs vital role in assessing the emotions and feeling of people as well as providing summaries of polarity results to organizations and other concerns. Reviews on products and personalities are very much important for organizations and personalities thus, text structurization divides the sentence properly to know the status of sentiments, actor or reviewer, reason of modification, place and time of modification. Structurization of text is done on the basis of five questions using five Ws. Text structurization of Sindhi text **شازيه فيس بڪ تي لکيو ته سامسنگ سٺي موبائيل فون آهي** (Shazia wrote on facebook that Samsung is good mobile phone) for sentiment summarization and analysis is done using five Ws questions properly.

**Who**: It shows that who expressed the opinion. For example **شازيه فيس بڪ تي لکيو** /**Who** (Shazia wrote on Facebook. /**Who**)

**What**: It shows the opinion of actor. For example Sindhi sentence **شازيه چيو ته سامسنگ سٺي موبائيل فون آهي** (Shazia said that Samsung is good mobile phone) is structured as:

شازيه چيو ته /**Who**

(DOI): 10.14569/IJACSA.2017.081038

سامسنگ سٺي موبائيل فون آهي /**What**

(Shazia said /Who Samsung is good phone /**What**.)

**Where**: It shows the place from where opinion is expressed or opinion is modified. For example Sindhi sentence **شازيه فيس بڪ تي لکيو ته سامسنگ سٺي موبائيل فون آهي** (Shazia Wrote on Facebook that Samsung is good mobile phone) is structured as:

شازيه /**Who**

فيس بڪ تي لکيو ته /**Where**

سامسنگ سٺي موبائيل فون آهي /**What**

(Shazia Wrote **/Who** on Facebook **/Where** that Samsung is good mobile phone **/What**.)

**Why**: It shows reason of modification of opinion. For example Sindhi sentence **شازيه فيس بڪ تي لکيو ته سامسنگ سٺي موبائيل فون آهي** is modified to compare Samsung mobile phone with others. The sentence is modified and structured as under:

شازيه پنهنجي راءِ کي/**Who**

فيس بڪ تي/**Where**

5 وڳين بدلائيندي لکيو ته /**When**

سامسنگ ٻين موبائيل فونز کان بهتر موبائيل فون آهي /**What**

(Shazia modified her opinion **/Who** in comment on Facebook /**Where** at 5 PM **/When** that Samsung is better Mobile phone than other mobile phones. /**What**.) The reason of modification in opinion is to compare Samsung mobile phone better than other mobile phones as it may be found better by opinionated person.

**When**: It shows the time of modification of opinion. When opinion is modified in above sentence than time is noticed. For example Shazia modified her opinion on Facebook regarding mobile phone at 5 PM.

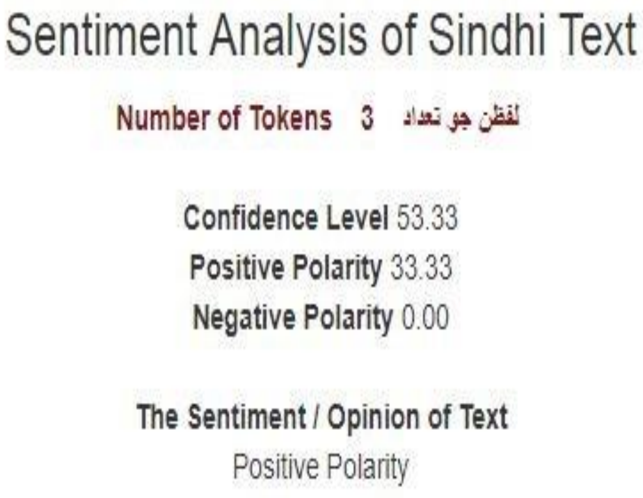
Thus, sentiment structurization is signification process for keeping opinion polarity records of users, which help the organizations, decision makers or concerned persons in knowing the current and previous opinions of users.

## III. Material and Methods

Problem of this research study is to evaluate the Sindhi text corpus for analysis of sentiment summerization and analysis. Sentiments show the view of people on different topics, thus structurization sections the text into separate topics. This study has tried to solve the NLP problems of Sindhi text sentiment analysis through structurization and machine learning supervised model. Model analyzed each part of structurization and sentiment polarity which are identified from Sindhi text corpus. Results show the performance of Sindhi NLP tool (http://www.sindhinlp.com/) for sentiment analysis and supervised classification model. Sindhi sentiment analysis has been done using Sindhi lexicons, which are identified through four part of speech like Noun, Adjective, Adverb and Verb.

Sindhi corpus is tagged with universal POS (UPOS) tag set to identify the senti-words. For exapmple Sindhi sentence **منهنجي ڪار سٺي آهي** (my car is good). Sindhi NLP tool tags this sentence like .AUX / ***آهي*** ADJ / ***سٺي*** NOUN / ***ڪار*** PRON / ***منهنجي*** In this sentence noun car is qualified by adjective good, which increase the confidence level of positive polarity. Sentiment analysis tool analyze the Sindhi text and finds out the senti-words. The senti-words are weighted with numbers. The weights are calculated to find out the average of each polarity. Finally, confidence level is measured on basis of high average rate of polarities which are positive and negative. High confidence level of polarity is described as the result of sentiment analysis.

Majority of sentiments are derived through Adjectives because adjectives qualify or disqualify the noun. The subjectivity of sentiment is found through adjectives as adjectives are very much important for sentiment analysis [10] (**Taboada** et al. 2011). For example Sindhi sentence **انب مٺو آهي** (Mango is sweet) presents the positive polarity. Polarity is assessed on basis of Sindh lexicon مٺو (Sweet) which is adjective whereas, Sindhi word **انب** (Mango) is noun. Sindhi word **سٺو** qualifies the Sindhi noun word **انب** Sindhi sentiment analysis tool has analyzed this sentence on basis of mapping UPOS tag set . The lexicon **مٺو** (Sweet) is tagged with UPOS ADJ which expresses positive sentiment thus the confidence level of positive polarity is high. Fig. 1 shows the sentiment analysis results of Sindhi **انب مٺو آهي** (Mango is sweet). Fig. 1 shows the number of lexicons which are used in the sentence, positive and negative weights as well as confidence level which shows average weight of both poalrities. As there is no negative polarity found in the sentence, thus confidence level is observed on basis of positive polarity only.



Sentiment Analysis of Sindhi Text

Number of Tokens 3 لفظن جو تعداد

Confidence Level 53.33

Positive Polarity 33.33

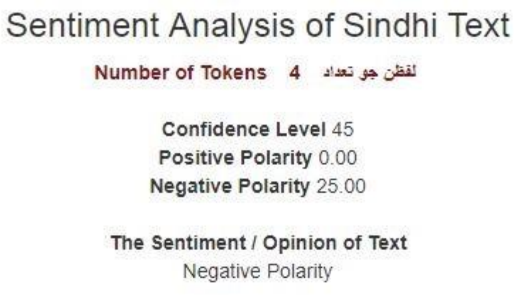
Negative Polarity 0.00

The Sentiment / Opinion of Text

Positive Polarity

Fig. 1. Sentiment analysis of Sindhi sentence showing positive polarity.

At the same time, another Sindhi sentence **هي انب ڪٽو آهي** (This Mango is sour) describes negative polarity of sentence. Here Sindhi adjective lexicon **ڪٽو** (sour) shows negative sentiment which leads to negative confidence level of polarity of sentence. Sindhi sentiment analysis tool observes tagging and polarity status of lexicons and performs sentiment analysis accordingly. Fig. 2 shows the sentiment analysis results of Sindhi sentence **هي انب ڪٽو آهي** (This Mango is sour).



Sentiment Analysis of Sindhi Text

Number of Tokens 4 لفظن جو تعداد

Confidence Level 45
Positive Polarity 0.00
Negative Polarity 25.00

The Sentiment / Opinion of Text
Negative Polarity

Fig. 2. Sentiment analysis of Sindhi sentence showing negative polarity.

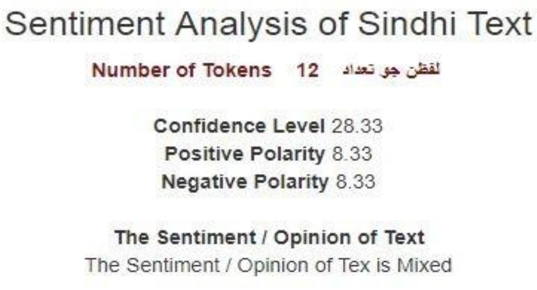
### A. Sindhi Corpus Dataset

Sindhi corpus is processed for sentiment structurization and analysis to build dataset for supervised machine learning processing. This dataset is developed for this research study, however, computational linguist may use this dataset for further research on Sindhi text and sentiment structurization analysis. This Dataset is comprised of 9779 records and 11 attributes. Polarity of sentences is identified with four categories which are shown in Table 1 with percentage of usage:

TABLE I. SINDHI CORPUS POLARITY IDENTIFICATION

| Polarity | Total Number in % |
|---|---|
| Positive | 60.32 |
| Negative | 11.15 |
| Mix | 7.55 |
| Neutral | 20.96 |

There is a large number of positive polarity and less number of mix polarity which are identified from Sindhi sentiment analysis dataset. Mix polarity is identified from those sentences which show both positive and negative polarities using discourse marker. Discourse marker separates the parts of Sentence. For example Sindhi sentence **اسان جو شهر سٺو آهي پر ماڻهو ان جو قدر نٿا ڪن** (Our city is good but people do not take care of it) presents two parts. First part **اسان جو شهر سٺو آهي** (Our city is good) which shows positive polarity because adjective lexicon **سٺو** (good) shows positive sentiment and second part **ماڻهو ان جو قدر نٿا ڪن** (people do not take care of it) shows negative polarity because Sindhi adverbial lexicon **نٿا** (not) shows negative sentiment. Sindhi discourse marker **پر** (but) connects both parts of sentence. Thus, polarity of this sentence is Mix. Fig. 3 shows the sentiment analysis results of Sindhi.



Sentiment Analysis of Sindhi Text

Number of Tokens 12 لفظن جو تعداد

Confidence Level 28.33
Positive Polarity 8.33
Negative Polarity 8.33

The Sentiment / Opinion of Text
The Sentiment / Opinion of Tex is Mixed

Fig. 3. Sentiment analysis of Sindhi sentence showing mixed opinion of text.

All records of corpus are structured with different number of Ws to segment the sentences. Each W takes dissimilar score from Sindhi corpus because all Ws are not used for each record. Table 2 shows total number of each W in percentage form. The **What** is used more than all other Ws because majority of sentences are showing opinions.

TABLE II. SINDHI CORPUS 5 WS STRUCTURIZATION

| Ws Structurization | Total Number (in %) of Ws annotated to Sindhi text |
|---|---|
| Who | 55.22 |
| What | 97.98 |
| Where | 42.07 |
| Why | 31.88 |
| When | 35.24 |

Sindhi corpus dataset is analyzed for identification of DTM and TF-IDF matrices using N-gram model, where N=3. The frequency of grams show the significance of Sindhi corpus dataset, thus, frequency is shown in form of document term matrix (DTM) and Term Frequency-Inverse Document Frequency (TF-IDF). DTM is consisted of C columns and D rows, therefore, M= C x D. Here, columns present the distinct language features which are vectors of matrix and rows show the number of documents which show the availability of features in documents. Sindhi language is complex language grammatically and morphologically, therefore, there is good number of adjoined words available in Sindhi corpus. Table 3 shows the distinct vectors and their frequency in Sindhi corpus dataset. DTM is comprised of 9779 rows × 2323 columns. Results of frequencies are total sum of Sindhi language features available in Sindhi corpus dataset. Results show the complexity of Sindhi corpus data set.

TABLE III. DTM OF SINDHI CORPUS DATASET USING TRI-GRAM MODE

| Frequency in all documents | Vectors / Features of DTM in tri-gram form |
|---|---|
| 18 | يا پيو ڪو |
| 18 | يا شعوري ڪوشش |
| 17 | هي خود شاگرد |
| 17 | هوندا آهن انقريب |
| 18 | هوشيار هجي امير |
| 17 | هن بدمست وڏيري |
| 17 | هائوس قومي ٽيم |
| 31 | ڳٺايو قوم فرض |
| 31 | ڳٺايو سنڌي ڳٺايو |
| ..... | ...... |
| 17 | آخر مرد عورت |
| 17 | آخر غربت بک |
| 31 | آخر سنڌ افغانين |
| 17 | آجياڻو نوازشريف خطاب |
| 17 | آباد ڪرين پٺائي |
| 11 | آباد ڪرين اتي |
| 70 | آءُ پنهنجو پسند |
| 18 | آءُ آڏو پيش |

Distinct terms are identified from the documents of Sindhi structured corpus through TF-IDF technique, thus, weight of features of corpus show the significance of terms. Study tracks the Latent Semantic Analysis (LSA) model to generate the TF-IDF matrix and analyze the relations of features with all documents available in Sindhi corpus data set. LSA extracts and conclude the relations between features and related documents automatically and statistically [11]. Stop words are removed from the Sindhi corpus dataset to build TF-IDF matrix. This matrix is two dimensional matrix, first dimension shows columns which show features and second dimension shows rows which present documents. Matrix is developed on basis of N-grams model where N=3. Table 4 shows the results of TF-IDF technique extracted from Sindhi corpus. TF-IDF matrix is comprised of 9779 documents and 4231 Sindhi language features.

Sindhi corpus dataset is significant dataset for supervised machine learning analysis. Sindhi language features are identified through DTM and TF-IDF to know the Sindhi language features and variations.

TABLE IV. TF-IDF OF SINDHI CORPUS DATASET USING TRI-GRAM MODEL

| Doc # | Feature Name | TF-IDF | Doc # | Feature Name | TF-IDF |
|---|---|---|---|---|---|
| **0** | 1860 | 0.4526 | 2 | 3369 | 0.2450 |
| **0** | 1068 | 0.5148 | 2 | 3004 | 0.2450 |
| **0** | 326 | 0.5148 | 2 | 482 | 0.2450 |
| **0** | 4175 | 0.5148 | 2 | 539 | 0.2304 |
| **1** | 4163 | 0.4009 | 2 | 1309 | 0.2304 |
| **1** | 1168 | 0.4580 | 2 | 3118 | 0.2450 |
| **1** | 4161 | 0.4580 | 2 | 3597 | 0.2450 |
| **1** | 1943 | 0.4580 | 2 | 2864 | 0.2450 |
| **1** | 4162 | 0.4580 | 2 | 3260 | 0.2304 |
| **...** | ... | ... | ... | ... | ... |
| **9773** | 143 | 0.2 | 9776 | 498 | 0.5329 |
| **9774** | 289 | 0.4082 | 9776 | 522 | 0.5983 |
| **9774** | 1933 | 0.4082 | 9776 | 1891 | 0.5983 |
| **9774** | 4062 | 0.4082 | 9777 | 507 | 0.3779 |
| **9775** | 3747 | 0.3731 | 9777 | 1491 | 0.3779 |
| **9775** | 264 | 0.4149 | 9778 | 290 | 0.7071 |
| **9775** | 2212 | 0.41491 | 9778 | 3048 | 0.7071 |

## IV. RESULT ANALYSIS

Supervised classification of Sindhi sentiment analysis corpus data set is done using dissimilar machine learning method to evaluate and assess the multi- classes. This study shows the comparative performance of supervised methods on Sindhi sentiment analysis corpus dara set. The performance of supervised classification methods is observed on basis of precision, recall and f-measure rates. F-measure combines the precision and recall; therefore, it is harmonic mean of precision and recall

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive + false positive}} \qquad (1)$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive + false negative}} \qquad (2)$$

The cross validation technique is used to validate and evaluate proper results. It works randomly using 10 folds to validate the training set which is 80% of data set and evaluate the test set which is 20% of data set. Structurization is analyzed using two machine learning methods. Structurization of Sindhi corpus is analyzed to know the precision, recall and f-measure scores of relevant records. Table 5 shows supervised machine learning analysis which is done on five Ws structurization. This analysis is performed using SVM Non-linear as dataset is labelled with multi-classes. Precision shows good number of positive predictive values of all Ws, whereas, sensitivity of all relevant records shows good recall results. Thus SVMS work better. Multiple hyper- planes divide all classes properly and give better results. Sindhi is complex language as it uses all forms of morphology as well as major number of bi-grams and tri-grams in its text. Therefore, results of SVM non-linear are better in this condition.

TABLE V. MACHINE LEARNING ANALYSIS OF 5 WS STRUCTURIZATION OF SINDHI DATASET USING SVMS CLASSIFIER

| Ws Structurization | Precision AVG | Recall AVG | F-score AVG |
|---|---|---|---|
| Who | 63 | 75 | 69 |
| What | 99 | 100 | 99 |
| Where | 61 | 62 | 60 |
| Why | 66 | 69 | 64 |
| When | 65 | 67 | 63 |

Another machine learning method K-NN is applied on Sindhi corpus dataset structurization to know the statistical results in shape of precision, recall and f-measure rates. K-NN is tested to know the proper value of K. Fig. 4 shows the better value of K for testing of accuracy of K-NN. In this study value of k is set to 2.

K-NN has analyzed all nearest neighbors according to value of K to evaluate and analyze the five Ws structurization. Results of precision, recall and f-measure are shown in Table 6. Precision shows better positive predicted values which are evaluated from retrieved records whereas, recall results show sensitivity of five Ws structurization which are recovered from all relevant records. F-score shows better accuracy of test data which is derived from Sindhi sentiment structurization corpus dataset. Thus, f-score validates the performance of binary classification which is done on Sindhi dataset.

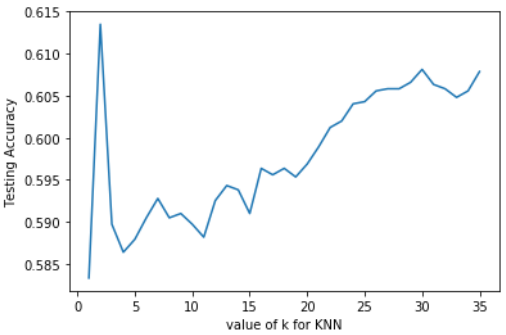


Fig. 4. Value of K for testing accuracy.

TABLE VI. MACHINE LEARNING ANALYSIS OF 5 WS STRUCTURIZATION OF SINDHI DATASET USING K-NN CLASSIFIER

| Ws Structurization | Precision AVG | Recall AVG | F-score AVG |
|---|---|---|---|
| Who | 65 | 62 | 63 |
| What | 99 | 100 | 99 |
| Where | 64 | 64 | 64 |
| Why | 67 | 69 | 68 |
| When | 65 | 66 | 65 |

Class polarity is labelled to know the accuracy rate of labelling of each target variable. This class is comprised of four polarity variables which are labelled to records according to their polarity. Both classifiers show different measurement rates. SVM non-linear classifier has performed better on Sindhi corpus dataset using cross validation with 10 folds. 80% of dataset is used as training dataset whereas remaining portion is used as test dataset. Measurement of dataset is observed through precision and recall scores. The relevant records are assessed with true relevant records through precision and relevant records are evaluated from all records through recall. However, results show better performance of supervised model which applied on Sindhi dataset. Table 7 shows measurement scores which are observed through true relevant and false relevant records from Sindhi data set using machine learning classifier SVMs.

TABLE VII. MACHINE LEARNING ANALYSIS OF SENTIMENT ANALYSIS OF SINDHI CORPUS DATASET USING SVMS

| Polarity | Precision AVG | Recall AVG | F-score AVG |
|---|---|---|---|
| Positive | 66 | 98 | 79 |
| Negative | 66 | 60 | 61 |
| Mix | 75 | 75 | 75 |
| Neutral | 89 | 22 | 35 |

Class polarity is also measured through K-NN classifier to differentiate the performance of both classifiers. Value of K is set to 2 to find out the nearest neighbors. Measurement of class polarity differentiates the assessment of true relevant and false relevant polarity categories. Table 8 shows the precision, recall and f-score of class polarity.

TABLE VIII. MACHINE LEARNING ANALYSIS OF SENTIMENT ANALYSIS OF SINDHI CORPUS DATASET USING K-NN

| Polarity | Precision AVG | Recall AVG | F-score AVG |
|---|---|---|---|
| Positive | 68 | 89 | 77 |
| Negative | 61 | 64 | 62 |
| Mix | 58 | 54 | 55 |
| Neutral | 42 | 21 | 28 |

## V. CONCLUSION

This study has tried to solve the NLP problems of Sindhi language as this language is one of the significant languages of Asia. There is no proper work done on sentiment analysis for Sindhi text thus, this is first work which is done on Sindhi language sentiment structurization and analysis. Sentiment structurization has solved the sentiment analysis problems of language for opinion tracking and sentiment summarization. Machine learning supervised analysis of own developed Sindhi sentiment structurization corpus dataset has proved the better performance of model. Precision, Recall, F-score and accuracy of supervised model has given good results on Sindhi corpus dataset. There is more need to work on Sindhi sentiment analysis for visual tracking that organizations can get proper opinions and reviews on their products.

## REFERENCES

[1] M. Abro, H. Nawaz, and W. Abro, “Performance Analysis of Dissimilar Classification Methods using RapidMiner,” Sindh Univ. Res., 2016.

[2] A. Ahmed and I. Elaraby, “Data Mining: A prediction for Student’s Performance Using Classification Method,” World J. Comput. Appl., 2014.

[3] D. Stephens and M. Diesing, “A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data,” PLoS One, 2014.

[4] I. Resnick, B. Verdine, and R. Golinkoff, “Geometric toys in the attic? A corpus analysis of early exposure to geometric shapes,” Early Child. Res., 2016.

[5] E. Cambria, B. Schuller, and Y. Xia, “New avenues in opinion mining and sentiment analysis,” IEEE Intell., 2013.

[6] S. Poria, E. Cambria, G. Winterstein, and G. Huang, “Sentic patterns: Dependency-based rules for concept-level sentiment analysis,” Knowledge-Based Syst., 2014.

[7] J. Mahar and G. Memon, “ALGORITHMS FOR SINDHI WORD SEGMENTATION USING LEXICON-DRIVEN APPROACH.,” Int. J., 2011.

[8] R. Motlani, “Developing language technology tools and resources for a resource-poor language: Sindhi.,” SRW@ HLT-NAACL, 2016.

[9] A. Das, S. Bandyaopadhyay, and B. Gambäck, “The 5w structure for sentiment summarization-visualization-tracking,” Process. Comput. …, 2012.

[10] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis,” Comput. Linguist., vol. 37, no. 2, pp. 267–307, Jun. 2011.

[11] C. Bosco, V. Patti, and A. Bolioli, “Developing corpora for sentiment analysis: The case of irony and senti-tut,” IEEE Intell. Syst., 2013.