Aspect-Based Sentiment Analysis of Sindhi Newspaper Articles

Irum Naz Sodhar[1*], Suriani Sulaiman[2], Abdul Hafeez Buller[3] and Anam Naz Sodhar[4]
irumnaz@sbbusba.edu.pk, ssuriani@iium.edu.my, ah.buller@quest.edu.pk anumakber10@gmail.com
[1] Post-Doctoral Fellow, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology
International Islamic University Malaysia.
[2]Assistant Professor, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology
International Islamic University Malaysia.
[3] Post-Doctoral Fellow, Department of Civil Engineering, Kulliyyah (Faculty) of Engineering International Islamic University Malaysia.
[4] Postgraduate Student, Quaid-e-awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan
[*]***Corresponding author:*** Irum Naz Sodhar, email: irumnaz@sbbusba.edu.pk

Summary

Aspect-Based Sentiment Analysis (ABSA) is a type of sentiment analysis that categorizes the polarity of sentiment based on the most important attributes of an entity related to a product, service or target. This research work presented a statistical study of Aspect-Based Sentiment Analysis on the dataset collected from the official website of Sindhi newspaper. There are various official newspaper websites available on the internet for Sindhi but focused on one official website that is Awami Awaz. In this research work, the dataset contains five sentences and one hundred and fifty-two words with punctuations, numbers and symbols chosen from the newspaper website. The dataset was first performed pre-processing task which were tokenized and after that identified Aspect-Based Sentiment Analysis from text context. The results of the Aspect-Based Sentiment Analysis are analyzed based one three categories: confidence level, positive polarity, and negative polarity. The pre-processing tasks were performed using the Sindhi NLP tool which is freely available online by a single search on google search engine. This tool contains multiple features available for the research purpose related to Sindhi Text.

Key words: *Aspect-Based Sentiment Analysis, Sindhi, sentiment analysis, polarity, aspects.*

1. Introduction

1.1 Sindhi Language

Sindhi language is widely used in Sindh-Pakistan as well as in different parts of the world. Majority of the people with Sindhi nationality live in Sindh and used Sindhi language as the official language of that province [1]. Nowadays, the use of the social media has been increasing rapidly as a platform for daily communication and other networking purposes. The public uses social networks groups for their communication and sharing their views, official documents, news and many others. The people residing in the Sindh province of Pakistan use social networks for their communications and other official/personal purposes in which the Sindhi script is used instead of English or other scripts [2].

Sindhi is one of the oldest languages in the world. It is morphologically rich with proper grammar and contains fifty-two (52) characters. Sindhi morphology is an expressed inner side of the Sindhi text. Morphological analysis of the Sindhi text involves characters, words, sentences, paragraphs, and documentations by different techniques [3]. Different sources of websites, forum platforms and newspaper articles provide information and different types of data utilization for different educational purposes. For our sentiment analysis, different types of techniques were used which includes machine learning techniques. These techniques are very helpful for the researchers and generate appropriate results for their solution [4-8].

1.2 Tool

Empirical study was conducted on the data collected from the official websites of Sindhi news articles known as awamiawaz. The aim of this research is to evaluate the aspect- based sentiment analysis of Sindhi texts. Sindhi has very limited resources as compared to English and other scripts [1][3]. The basic tasks of NLP (Natural Language Processing) involve the analysis of the nature of texts which includes lemmatization, stemming, parsing, parts-of-speech tagging, spell checking, information retrieval, machine transliteration and many more. In this this research work, the dataset was collected from the official website of Sindhi news and Aspect-based Sentiment Analysis (ABSA) is performed by using the Sindhi NLP online tool [9]. A small number of research works on Sindhi text have been conducted in the tasks of word tokenization, part-of-speech tagging and sentiment analysis [10].

2. Materials and Methods

2.1 Sindhi Script

Sindhi is a script with fifty-two (52) characters for writing as well as for speaking. It is one of the oldest

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.54

scripts amongst the languages and it has its own grammatical rules for writing and speaking. Sindhi language is rich by morphological structure. This language has two gender types for nouns, which are: masculine and feminine [11]. Figure 1 illustrates an example of nouns with the Masculine (M) and Feminine (F) genders.

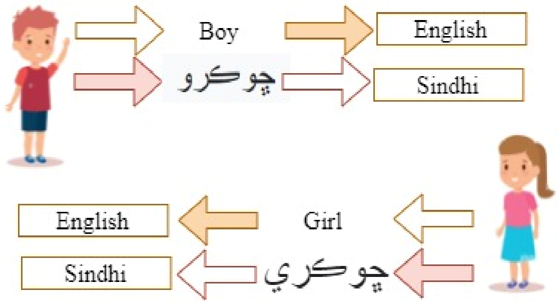


Fig. 1. Masculine and Feminine genders (English -Sindhi)

2.2 Computational Linguistics (CL) Tool for Sindhi Language

The Computational Linguistics Tool for Sindhi Language [12] is an NLP tool widely used for the analysis of Sindhi scripts. It provides solutions for many computational problems related to Sindhi scripts. Some of the latest features available in this tool include Sindhi Online Parser, Sindhi WordNet, Sindhi Lemma, Sindhi Stemmer, Sentiment Analysis and Aspect-based Sentiment Analysis (ABSA)[9]. As one of the ancient scripts of the world with fifty-two (52) characters, the dot character is used for the appropriate appearance of the letters. The dots or periods (.) are categorized into four different groups namely the single dot characters, double dots character, triple dots characters and four dots character as used in Sindhi script [13]. Sindhi script is a right-handed script similar to Arabic and Urdu scripts [14, 15].

2.3 Features of Sindhi NLP

Sindhi Natural Language Processing (SNLP) is dealing with Sindhi script and gives computational solution of problems occurring in Sindhi script. Mostly researcher focus on the English tokenization of words or letters, part-of-speech (POS) tagging, parsing, sentiment analysis, Aspect-based Sentiment Analysis (ABSA) and text summarization [16]. To date, various online tools are available for English scripts while for Sindhi, the number of online tools available is very limited. Developers for Sindhi NLP tools mainly focus on text parser of Sindhi script, Sindhi dataset analysis, statistical result analysis, word tokenization, POS Tagging and Aspect based Sentiment Analysis which are very helpful for researchers working with the Sindhi language [17, 18].

i. Parser

A text parser expresses the arrangement of texts for the scripts. Sindhi CL tool performs NLP task only on Sindhi scripts. This tool produces analysis results for Sindhi script within a few seconds. Sindhi online parser uses the Universal Part-of-Speech (UPOS) Tagging and Sindhi Parts-of-Speech (SPOS) tagging to tag and syntactically parse the Sindhi dataset. The statistical results display the execution time, number of word tokens, frequencies of phrases, as well as the UPOS and SPOS morphological forms of Sindhi texts.

ii. WordNet of Sindhi

WordNet of Sindhi is a database/knowledge-based tool for Sindhi scripts. It comprises of a group of data in the form of proverbs, poetry, social network comments, news and essays. The relations between text are sometimes difficult to understand when they have different senses in which the WordNet helps to solve this problem.

iii. Lemmatization and Stemming

Lemmatization is the process of finding the lemma of a word and identifies its part-of-speech. Lemma is the basic form of word which makes sense of understanding. Stemming gives you an idea about word without knowledge of the contexts. In a variety of scripts, the text is represented in different forms. For example, in English, ‘to learn’ may possibly become visible as ‘learn’ and ‘learnt’ while the word ‘learning’ has the base form ‘learn’ in the dictionary. Thus, ‘learn’ is the lemma for all the words mentioned above.

3. Process of Evaluation

The evaluation process depends on multiple components and tasks involved in the aspect-based sentiment analysis which comprises of the dataset collected, the input data, the pre-processing of the data and the extraction process of the aspect-based word as shown in Figure 2.

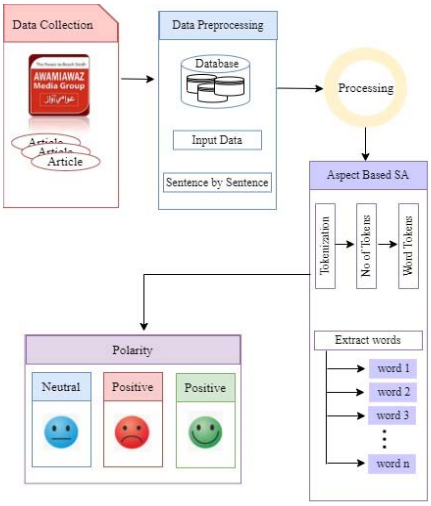


Fig. 2. Process of Evaluation

3.1 Data Collection

Variety of Sindhi data is available on different internet platforms such as the Sindhi blogs, online forum, websites, magazines and social media. In this research, we collected data from the official website of online Sindhi newspaper [19]. This dataset contains a total of five sentences and one hundred and fifty-five words with punctuations and symbols as shown in Table 1.

Table 1 Data set of Aspect Based Sentiment Analysis

Sentences	S.NO
طبي ماهرن اڏ رنگ ۽ دل جي دوري جهڙن موت مار مرضن بلڊ پريشر کان بچڻ جو طريقو ٻڌائي ڇڏيو.جو سبب ٻڻجندڙ هاءِ	01
چين ۾ ٿيل هڪ طبقي تحقيق ۾ هاءِ بلڊ پريشر کان بچڻ لاءِ آنن، چڪن ۽ پروٽين سان ڀرپور هر خوراڪ کي پنهنجي غذا جو لازمي حصو بڻائڻ جو چيو ويو آهي	02
سدرن ميڊيڪل يونيورسٽي جي هن تحقيق ۾ ٻڌايو ويو آهي ته آنا، چڪن، ٻج، سي فوڊ ۽ گوشت جهڙين وڌيڪ پروٽين وارين غذائن جو استعمال هاءِ بلڊ پريشر جو خطرو گهٽائڻ ۾ مددگار ثي سگهي ٿو	03
هن تحقيق ۾ 12 هزار کان وڌيڪ چيني نوجوانن جي جاچ گڏ ڪري چندڊ چاڻ ڪئي وئي هئي، جنهن ۾ انهن جي غذائي عادتن ۽ بلڊ پريشر جي پيٽ 6 سالن تائين ڪئي وئي	04
هنن ماڻهن کي پروٽين جيحاصل ڪرڻ وارن ذريعن جي استعمال جي بنياد تي اسڪور ڏنا ويا ۽ ان مقصد لاءِ پاڻ مرادو رپورٽ ڪيل سرويز جو سهارو ورتو ويو	05

3.2 Input data

In this research, the inputs are in the form of five sentences which are pre-processed before being analyzed. The first sentence contains twenty-five words with one punctuation, the second sentence contains thirty-two words with no punctuation, the third sentence contains thirty-six words with three punctuations, the fourth sentence contains thirty-four words with a single punctuation (i.e., comma) and the fifth sentence contains twenty-nine words without any punctuations.

3.3 Data Processing

In this research, we used Sindhi CL tool to process and analyze the data. Sindhi CL tool is easily available online for use in empirical studies of Sindhi texts. Sindhi texts data were collected from official website of Sindhi newspaper. One hundred fifty-five words were process in this tool to identify aspect based sentiment analysis. This text includes sentences with words, punctuation and symbols. Sindhi text data analyzed in online available Sindhi CL tool and it gave the results.

3.3.1 Tokenization

Word tokenization is another feature available in the online Sindhi CL tool. This online tool returns the analysis of a large amount of Sindhi text within a few seconds. This tool tokenized the data word by word and yields the total number of tokens in the texts.

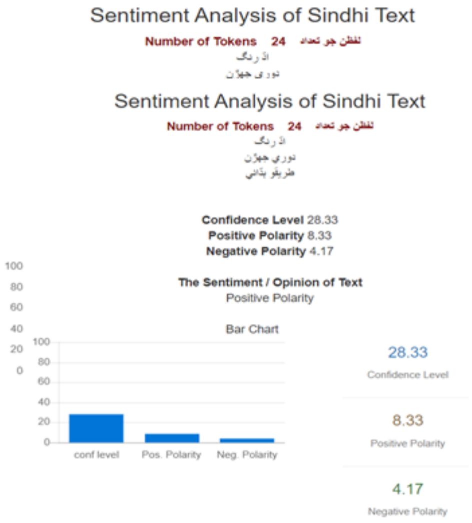
3.4 Aspect-Based Words

Traditional file-level sentiment categorization tries to pick out the overall sentiment polarity of a given textual content as positive, negative and confidence or neutral. Unlike whole document level sentiment, aspect-level sentiment type classifies the sentiment of one specific text in its context sentence. In this research, the Sindhi CL tool identified the aspect-based words from the input text and returns the total number of word tokens and the aspect-based words.

4. Results and Discussion

4.1 Sentence by sentence

Result of sentence one shows the total number of tokens (twenty-24), Aspect based Sentiment Analysis comes six words and confidence level percentage is 28.33, positive polarity 8.33, negative polarity 4.17 and Overall sentiment comes positive polarity shown in figure (a). Result of sentence two shows total number of tokens is thirty two (32), Aspect based Sentiment Analysis come two words and confidence level percentage is 26.25, positive polarity 6.25, negative polarity 0.00 and Overall sentiment comes positive polarity shown in figure (b).Result of sentence three shows total number of tokens (Thirty six-36), Aspect based Sentiment Analysis comes four words and confidence level percentage is 20, positive polarity 0.00, negative polarity 0.00 and Overall sentiment comes neutral polarity shown in figure (c). Result of sentence four shows total number of tokens (Thirty four-34), Aspect based Sentiment Analysis comes two words and confidence level percentage is 25.88, positive polarity 5.88, negative polarity 0.00 and Overall sentiment comes positive polarity shown in figure (d).Result of sentence fifth shows total number of tokens (Twenty nine-29), Aspect based Sentiment Analysis comes no any word and confidence level percentage is 26.9, positive polarity 6.90, negative polarity 0.00 and Overall sentiment comes positive polarity shown in figure (e).



Sentiment Analysis of Sindhi Text

Number of Tokens 24 لفظن جو تعداد

اڏ رنگ

نوري جهڙن

Sentiment Analysis of Sindhi Text

Number of Tokens 24 لفظن جو تعداد

اڏ رنگ

نوري جهڙن

طريقو بڻائي

Confidence Level 28.33
Positive Polarity 8.33
Negative Polarity 4.17

The Sentiment / Opinion of Text
Positive Polarity

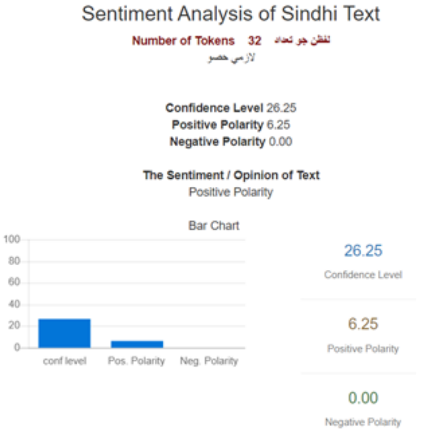
Bar Chart

28.33
Confidence Level

8.33
Positive Polarity

4.17
Negative Polarity

Fig a. Result of Sentence 1



Sentiment Analysis of Sindhi Text

Number of Tokens 32 لفظن جو تعداد

لازمي حصو

Confidence Level 26.25
Positive Polarity 6.25
Negative Polarity 0.00

The Sentiment / Opinion of Text
Positive Polarity

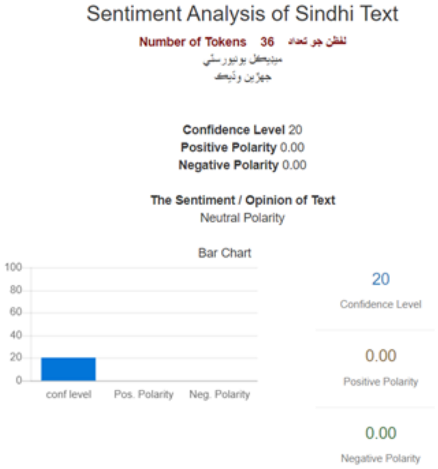
Bar Chart

26.25
Confidence Level

6.25
Positive Polarity

0.00
Negative Polarity

Fig b. Result of Sentence 2



Sentiment Analysis of Sindhi Text

Number of Tokens 36 لفظن جو تعداد

ميڊيڪل يونيورسٽي

جهڙين وٽيڪ

Confidence Level 20
Positive Polarity 0.00
Negative Polarity 0.00

The Sentiment / Opinion of Text
Neutral Polarity

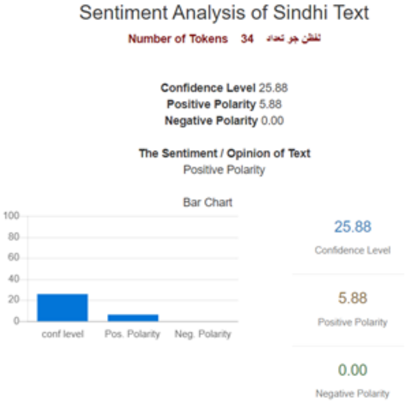
Bar Chart

20
Confidence Level

0.00
Positive Polarity

0.00
Negative Polarity

Fig c. Result of Sentence 3



Sentiment Analysis of Sindhi Text

Number of Tokens 34 لفظن جو تعداد

Confidence Level 25.88
Positive Polarity 5.88
Negative Polarity 0.00

The Sentiment / Opinion of Text
Positive Polarity

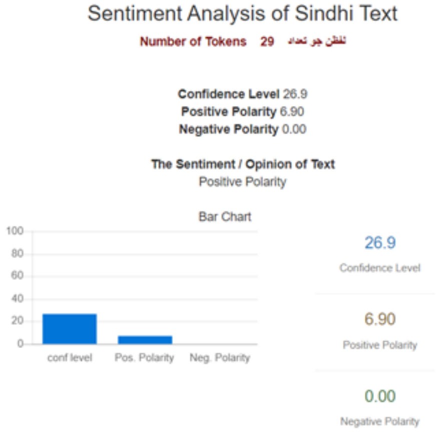
Bar Chart

25.88
Confidence Level

5.88
Positive Polarity

0.00
Negative Polarity

Fig d. Result of Sentence 4



Sentiment Analysis of Sindhi Text

Number of Tokens 29 لفظن جو تعداد

Confidence Level 26.9
Positive Polarity 6.90
Negative Polarity 0.00

The Sentiment / Opinion of Text
Positive Polarity

Bar Chart

26.9
Confidence Level

6.90
Positive Polarity

0.00
Negative Polarity

Fig e. Result of Sentence 5

Fig. 3 Sentiment Analysis of Sentences (1-5)

Table 2 Overall Results on the Basis of ABSA on Tool

Sentences	***Total No. of Words (W)***	***Total No of Dots (D)***	***Total No of Symbols (S)***	***Total No. of words ABSA***	***ABSA Words***
One	(24)	(69)	(03)	(06)	اڌ رنگ دوري جھڙن طريقو ٻڌائي
Two	(32)	(94)	(07)	(02)	لازمي حصو
Three	(36)	(118)	(06)	(04)	ميڊيڪل يونيورسٽي جھڙين وڌيڪ
Four	(34)	(103)	(08)	(02)	پيٽ سالن
Five	(29)	(68)	(04)	---	---

Problematic Equations

From the above table below equation are generated

$$\text{Sentence} = W \pm \text{ABSA} \tag{1}$$

$$WT = (DT + ST) \tag{2}$$

$$\text{ABSA} = \left\{ \begin{array}{l} \Sigma(Pw), \\ \Sigma(Negw) \\ \Sigma(Neuw) \end{array} \right\} \tag{3}$$

$$\Sigma(Pw) = (Pw1+Pw2+Pw3+\ldots Pwn) \tag{a}$$

$$\Sigma(Negw) = (Negw1+Negw2+Negw3+\ldots Negwn) \tag{b}$$

$$\Sigma(Neuw) = (Neuw1+Neuw2+Neuw3+\ldots Neuwn) \tag{c}$$

$$\text{Polarity of Sentence (Positive)} = (\Sigma(Pw) \div \text{ABSA}) \tag{4}$$

$$\text{Polarity of Sentence (Negative)} = (\Sigma(Negw) \div \text{ABSA}) \tag{5}$$

$$\text{Polarity of Sentence (Neutral)} = (\Sigma(Neuw) \div \text{ABSA}) \tag{6}$$

Whereas:

WT	=	Total No. of Words
ABSA	=	Aspect Based Sentiment Analysis
DT	=	Total No of Dots
ST	=	Total No of symbols
Pw	=	Positive words
Negw	=	Negative words
Neuw	=	Neutral words

Table 2 Polarity of sentences

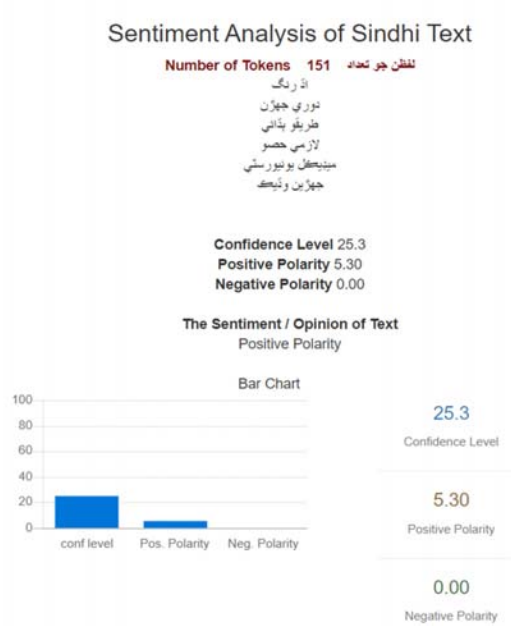
Sentences	***Results of Levels of Polarity Sentence by Sentence (%)***			***Overall Sentiment / Opinion of Text***
	Confidence Level	***Positive Polarity***	***Negative Polarity***	
1	28.33	8.33	4.17	Positive Polarity
2	26.25	6.25	0.00	Positive Polarity
3	20	0.00	0.00	Neutral Polarity
4	25.88	5.88	0.00	Positive Polarity
5	26.9	6.90	0.00	Positive Polarity

The accuracy of the tokenization results using the Sindhi CL tool when used with the complete dataset is 97.41%. The accuracy of the tokenization results produced by the Sindhi CL tool when the input used is sentence by sentence is 100%.

5. Sentiment Analysis

In the Sindhi CL tool, sentiments are categorized into three categories which are confidence level, positive polarity and negative polarity. Confidence level shows the percentage of input data which are neutral, positive polarity shows the percentage input with positive sentiment and negative polarity shows the percentage of how much of the input data have negative sentiments. The overall result produced by this tool showed three categories of sentiments with the bar chart for each category of sentiment.

A statistical result of Sindhi text of newspaper shows the number of tokens, Aspect based words and Categories of Sentiments with percentage, Overall Sentiment Analysis of text and Bar Chart of Sentiment categories with percentages. In this research study Results evaluates on tool into two categories such as: one is complete dataset and other one is sentence by sentence. Overall result of whole data set shows number of tokens (one hundred fifty one-151), Aspect based Sentiment Analysis comes twelve words and confidence level percentage is 25.3, positive polarity 5.30, negative polarity 0.00 and Overall sentiment comes positive polarity shown in figure 4.



Sentiment Analysis of Sindhi Text

Number of Tokens 151 لفظن جو تعداد

اڏ رنگ
ٽوري جهڙن
طريقو ٻڌائي
لازمي حصو
ميڊيڪل يونيورسٽي
جهڙين وڌيڪ

Confidence Level 25.3
Positive Polarity 5.30
Negative Polarity 0.00

The Sentiment / Opinion of Text
Positive Polarity

Fig. 4. Complete Dataset Result

6. Conclusion and Future work

This research work described the Aspect-based Sentiment Analysis (ABSA) using the dataset from Sindhi official website of news articles focusing on two NLP tasks mainly tokenization and sentiment analysis. The NLP tasks were performed using the Computational Linguistics Tools for Sindhi Language which is freely available on internet to perform common NLP tasks for Sindhi language using Sindhi dataset. The tool produced an overall accuracy of 97.41% with four Positive Polarity and one Neutral Polarity for the Aspect-based Sentiment Analysis (ABSA) on five (5) Sindhi sentences. For future work, to use this tool on a larger dataset of Sindhi texts in different domains.

References

[1] J. A. Mahar, G. Q. Memon. Rule based part of speech tagging of Sindhi language. In 2010 International Conference on Signal Acquisition and Processing 2010 Feb 9, pp. 101-106. IEEE.

[2] D. Wang, M. Fang, Song Y, Li J. Bridging the gap: Improve part-of-speech tagging for Chinese social media texts with foreign words. In Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5) 2019, pp. 12-20.

[3] R. Motlani, F. Tyers, D. Sharma. A Finite-State Morphological Analyser for Sindhi. In Proceedings of the Tenth International Conference on

[4] Language Resources and Evaluation (LREC 2016), pp:2572-2577.

[5] M. E. Tibbo, C. C. Wyles, S. Fu, S. Sohn, D. G. Lewallen, D. J. Berry, H. M. Kremers. Use of natural language processing tools to identify and classify per prosthetic femur fractures. The Journal of arthroplasty. 2019 Oct 1;34(10):2216-9.

[6] Sodhar, I. N., Bhanbhro, H., & Amur, Z. H. (2019). Evaluation of web accessibility of engineering university websites of Pakistan through online tools. IJCSNS, 19(12), 85-90.

[7] Khairnar, J., & Kinikar, M. (2015). Sentiment analysis based mining and summarizing using SVM-MapReduce. International Journal of Computer Science and Network Security (IJCSNS), 15(5), 85.

[8] Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), 62.

[9] N. X. Bach, N. D. Linh, T. M. Phuong. An empirical study on POS tagging for Vietnamese social media text. Computer Speech & Language. 2018 Jul 1;50:1-5.

[10] I. N. Sodhar, A. H.Jalbani, M. I. Channa, D. N. Hakro. Parts of Speech Tagging of Romanized Sindhi Text by applying Rule Based Model. IJCSNS. 2019 Nov;19(11), pp:91.

[11] M. Ali, A. I. Wagan. An Analysis of Sindhi Annotated Corpus using Supervised Machine Learning Methods. Mehran University Research Journal ofEngineering and Technology. 2019 Jan 1;38(1):185-96.

[12] M. A. Dootio, A. I. Wagan. Syntactic parsing and supervised analysis of Sindhi text. Journal of King Saud University-Computer and Information Sciences.2019 Jan 1;31(1):105-12.

[13] M. A. Dootio. Computational Linguistic Tools for SindhiLanguage.https://sindhinlp.com/sentimentABSA.php

[14] I. N. Sodhar, A. H.Jalbani, M. I. Channa, D. N. Hakro. Identification of Issues and Challenges in Romanized Sindhi Text. . International Journal of Advanced Computer Science and Applications (IJACSA), 2019 Sep;10(9): 229-233.

[15] S. S. Rizvi, A. Sagheer, K. Adnan, A. Muhammad. Optical Character Recognition System for Nastalique Urdu-Like Script Languages using Supervised Learning. International Journal of Pattern Recognition and Artificial Intelligence. 2019 Mar 6:1953004.

[16] J. A. Mahar, G. Q. Memon. Sindhi part of speech tagging system using WordNet. International Journal of Computer Theory and Engineering. 2010 Aug 1;2(4):538.

[17] Ahmed, S., Hina, S., Atwell, E., & Ahmed, F. (2017). Aspect based sentiment analysis framework using data from social media network. International Journal of Computer Science and Network Security, 17(7), 100-105.

[18] Khairnar, J., & Kinikar, M. (2015). Sentiment analysis based mining and summarizing using SVM-MapReduce. International Journal of Computer Science and Network Security (IJCSNS), 15(5), 85.

[19] Daily Awami Awaz. Online Sindhi newspaper of dated 11th March, 2022.

[20] Sodhar, I. N., Buller, A. H., & Sodhar, A. N. (2021). Identification of Online Statistical Translation and Text Issues in Communication Technologies. International Journal of Advanced Trends in Computer Science and Engineering, 10(2).

[21] Sodhar, I. N., Bhanbhro, H., Amur, Z. H., Jalbani, A. H., & Buller, A. H. Sindhi Language Processing on Online SindhiNLP Tool. vol, 4, 4-7.

[22] Sodhar, I. N., Jalbani, A. H., Buller, A. H., & Sodhar, A. N. (2020). Tools Used In Online Teaching and Learning through Lock-Down. International Journal of Computer Engineering in Research Trends, no, 8, 36-40.

[23] Khairnar, J., & Kinikar, M. (2015). Sentiment analysis based mining and summarizing using SVM-MapReduce. International Journal of Computer Science and Network Security (IJCSNS), 15(5), 85.

[24] Daily Awami Awaz. Online Sindhi newspaper of dated 11th March, 2022.