Mehran University Research Journal of Engineering & Technology
Vol. 38, No. 1, 185-196, January 2019
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1901.15

An Analysis of Sindhi Annotated Corpus using Supervised Machine Learning Methods

MAZHAR ALI*, AND ASIM IMDAD WAGAN**

RECEIVED ON 14.06.2017 ACCEPTED ON 25.05.2018

ABSTRACT

The linguistic corpus of Sindhi language is significant for computational linguistics process, machine learning process, language features identification and analysis, semantic and sentiment analysis, information retrieval and so on. There is little computational linguistics work done on Sindhi text whereas, English, Arabic, Urdu and some other languages are fully resourced computationally. The grammar and morphemes of these languages are analyzed properly using dissimilar machine learning methods. The development and research work regarding computational linguistics are in progress on Sindhi language at this time. This study is planned to develop the Sindhi annotated corpus using universal POS (Part of Speech) tag set and Sindhi POS tag set for the purpose of language features and variation analysis. The features are extracted using TF-IDF (Term Frequency and Inverse Document Frequency) technique. The supervised machine learning model is developed to assess the annotated corpus to know the grammatical annotation of Sindhi language. The model is trained with 80% of annotated corpus and tested with 20% of test set. The cross-validation technique with 10-folds is utilized to evaluate and validate the model. The results of model show the better performance of model as well as confirm the proper annotation to Sindhi corpus. This study described a number of research gaps to work more on topic modeling, language variation, sentiment and semantic analysis of Sindhi language.

Key Words: Machine Learning, Sindhi Corpus, Universal part of speech, Random Forest, Support Vector Machines, Natural Language Processing.

1. INTRODUCTION

Every person, having an internet connection, can utilize the international resources for education, health, social development and etc. Websites, blogs and online social forums provide all types of data, therefore, to manage and analyze that data, is a problem of real world. Machine learning solves these problems through its dissimilar techniques. Classification is a noteworthy technique of machine learning and data mining [1], which predicts the results on the basis of training data set [2]. Now-a-days, there are several websites, blogs and social media sources which produce a large number of data.

This study, has developed the plain corpus by collecting data through internet resources and annotated that plain

Authors E-Mail: (mazharaliabro@gmail.com, aiwagan@gmail.com)
* Benazir Bhutto Shaheed University, Lyari, Karachi, Pakistan.
** Mohammad Ali Jinnah University, Karachi, Pakistan.

corpus with universal POS tag set and Sindhi POS tag set using online NLP (Natural Language Processing) resource (www.sindhinlp.com). We have analyzed Sindhi annotated corpus through machine learning model which is consisted of two supervised machine learning methods: (1) SVM (Support Vector Machine) non-linear and (2) RF (Random Forest). Purpose of this research study is to present the grammatical and morphological variation of Sindhi. Sindhi is a less resourced language [3,4] in comparison of English language. Nevertheless, some work has been done on tokenization and POS tagging of Sindhi text [5-7] as well as NLP tools are accessible online for solution of Sindhi linguistic problems [7]. In this connection, Sindhi Devanagari script [8] for POS tagging system is not helpful for right hand written script of Sindhi text. NLTK (https://www.nltk.org/) is one of the prominent resources for solution of linguistics problems of human languages but it does not support fully to Sindhi corpus analysis process because of unavailability of libraries for Sindhi stemming and morphemes, which are different than the English stemming words and morphemes. Therefore, this study has developed its own program to analyze the Sindhi annotated corpus.

1.1 Sindhi Language

Sindhi is a complete language having a culture, land, civilization, history, proper grammar and rich morphological structure with fifty-two alphabets [9]. Sindhi is an indigenous language having all properties of native and complete language [10]. Sindhi script is one of the oldest scripts of the world having all required features. Sindhi text is rich enough with its literary and non-literary text material. The grammar of Sindhi language is quite different from the majority of languages of the world. Therefore, computational linguistics analysis process may not be the same for Sindhi language as for English or other languages of the world. Grammatically, gender in Sindhi language is of two types, which are masculine (جنس مذڪر) and feminine (جنس مونث). Those nouns which show meaning with male (نر) is called a masculine gender and nouns which show meaning with female (مادي) is called a feminine gender. These sub-classes of noun correspond to pronouns, adjectives, verbs and determiners. All types of nouns, adjectives, pronouns, verbs and determiners come within these two genders. Sindhi is a comprehensive language, therefore, there are different names or pronunciations for other genders. For example, chhokro چوڪرو (a boy), chhokri چوڪري (a girl), haathi هاٿي (an elephant), haathinn هاٿڻ (a Female elephant). The word 'Teacher' is a common noun, therefore, it is same for male and female in English language whereas, it is pronounced in a different way for male teacher and female teacher in Sindhi language. Male teacher is pronounced as Ustaad (استاد) and female teacher is pronounced as Ustaadiani or Ustade (استاديائي or استاڻي) in Sindhi language.

Compound verbs make Sindhi language more beautiful and complex. Single inflection or diacritic and addition of single character changes the complete meaning of compound verbs. Sindhi compound verbs اچي ٿو (Achay tho) and اچيو ٿو (Achev tho) are totally different from each other according to their lexical or contextual meaning. First compound verb shows that someone comes and second compound verb shows that someone comes to you. Compound verbs may be connected with nouns, verbs and adjective. To identify Sindhi compound verbs for annotation is problematic for Sindhi tagger and corpus development tools.

Verb is a very important part of speech for any language because it constructs the sentence properly. Almost, all types of verb are same in Sindhi and other languages except intransitive verb. Practically, all languages use intransitive verb in active voice but no language uses intransitive verb in passive voice except Sindhi language [11]. This is the uniqueness of Sindhi language that it uses intransitive verb in passive voice. The example of active and passive voice verbs is shown with a simple sentence of Sindhi language: آءُ پَٽ تي سمهان ٿو (I sleep on earth). Intransitive passive voice of presented Sindhi simple sentence is: پَٽ تي سمهجي ٿو. (pat te sumhjay tho). Thus, these types of the differences make the Sindhi as a unique and significant language of the world.

1.2 Corpus Annotation Process

Sindhi annotated corpus is tagged with Universal POS and Sindhi POS tag sets. Both tag sets are important for Sindhi corpus annotation process. The assessment and study of tree banks presents the rank of UPOS (Universal Part-of-Speech) tag set for dissimilar languages of the world because tag sets are language specific [12]. There is a similarity between Sindhi POS tag set and UPOS tag set with a little difference while majority of UPOS tags are similar to SPOS tag set. Sindhi NLP tool http://www.sindhinlp.com/ tokenizes the Sindhi text documents into separate tokens, described in Fig. 1 and annotates them with universal part of speech, described in Fig. 2 and with Sindhi part of speech, described in Fig. 3.

The tokenization and annotation processes stand the Sindhi token as complete token for understanding and analysis. Table 1 shows the resemblance of Sindhi POS tag set with universal POS Tag set. The UPOS tag X is used for those words which are unknown or cannot be identified according to universal POS tag set. UPOS article PART is used for negation showing words and possession marker while Sindhi POS Adverb is used for negation showing words and possessive case or pronoun is used for possession marker.

Mostly, it is observed that Penn tree bank tags [13] or UPOS tags are used to annotate corpus of any language, but this study maps Sindhi POS tag set along with Universal POS tag set to Sindhi corpus which shows the significance of Sindhi POS tag set. Table 2 shows

Tokenization

”سنڌي“-1،”ٻولي“-2،”دنيا“-3، ”جي“-4، ”پراڻي“-5،”ٻولي“-6، ”آهي“-7،

FIG. 1. TOKENIZATION OF SINDHI TEXT DOCUMENT

UPOS Tagging

سنڌي/PROPN ٻولي/NOUN دنيا/NOUN جي/ADP پراڻي/ADJ ٻولي/NOUN آهي/AUX

FIG. 2. ANNOTATION OF SINDHI TEXT DOCUMENT WITH UNIVERSAL POS TAG SET

سنڌي نشان

سنڌي/اسم خاص ٻولي/اسم دنيا/اسم جي/حرفِ جر پراڻي/صفت ٻولي/اسم آهي/فعل معاون)

FIG. 3. ANNOTATION OF SINDHI TEXT DOCUMENT WITH SINDHI POS TAG SET

frequency of UPOS and SPOS tags which are measured during the annotation process to Sindhi corpus. Dissimilarity appears in UPOS tags PART, Adverb and SPOS tag Adverb because this study does not use PART in place of Sindhi Adverb and possessive marker or pronoun. Thus, number of frequencies of Adverb of SPOS is high than the number of frequencies of Adverb of UPOS.

The most common word which is used frequently in Sindhi corpus is جي(of). This word is preposition (حرفِ جر) in Sindhi language. جي (jay) lexicon of Sindhi language describes the relation, correspondence or dependency in Sindhi text. The unique or frequently used words are text vectors in document term matrix. Table 3 shows the detail of total words, top words and frequently used word in Sindhi corpus.

TABLE 1. SIMILARITY BETWEEN UPOS AND SPOS TAG SETS

Universal POS Tag	Sindhi POS Tag	Universal POS Tag	Sindhi POS TAg
ADJ	صفت	ADP	حرفِ جر
ADV	ظرف	AUX	فعل معاون
CONJ	حرف جملو	DET	ضمير اشارو
INTJ	حرف ندا	NOUN	اسم
NUM	عددي صفت	PART	ضرف، حرف اضافت، ضمير
PRON	ضمير	PROPN	اسم خاص
PUNCT	بيهڪ جي نشاني	SCON	حرف جملو شرطيه
SYM	نشانيون	VERB	فعل
X	نامعلوم		

TABLE 2. FREQUENCY OF UNIVERSAL AND SINDHI POS TAGS

UPOS	Frequency in Dataset	SPOS	Frequency in Dataset
NOUN	1080	اسم	1080
PROPN	131	اسم خاص	131
PRON	135	ضمير	135
DET	172	ضمير اشارو	172
VERB	438	فعل	438
AUX	91	فعل معاون	91
ADJ	358	صفت	358
NUM	65	صفت عددي	65
ADV	169	ظرف	204
CONJ	181	حرف جملو	181
ADP	667	حرفِ جر	667
INTJ	1	حرف ندا	1
PART	35		
X	30	نامعلوم	30

2. MATERIALS AND METHOD

This study developed a new dataset for Sindhi language with the intention of creating a standardized corpus for NLP researchers. Language features were extracted through TF-IDF model using n-gram model where N = 1. TF-IDF is arrangement of two parts which are "Term frequency" and "Inverse term frequency". TF sums the frequency of words available in corpus and IDF describes the word with all its information. Therefore, IDF is measured as the logarithm of documents available in corpus. Equations (1-2) shows the working procedure of TF-IDF.

$$f_{t,d} / \sum_{t' \in d} f_{t',d} \sum_{t' \in d} f_{t',d} \quad (1)$$

$$\mathrm{idf}(t, D) = \log \frac{N}{\left|\{d \in D : t \in d\}\right|} \quad (2)$$

TF-IDF finds significant words/feature names from Sindhi corpus which perform important role in documents. Feature names are special terms which are deferent from each other and significant for corpus documents. Table 4 shows feature names, extracted from Sindhi corpus using n-gram model where n=1.

2.1 Sindhi Annotated Corpus

The Sindhi annotated corpus is multi-class and multi-featured corpus dataset. There are four features of Sindhi annotated data set: (1) UPOS tagging (2) SPOS tagging (3) Lemmatization and (4) stemming. These all features are vital for Sindhi corpus analysis because UPOS and SPOS tagging of Sindhi corpus facilitate in syntactic, semantic and sentiment analysis. Lemmatization process stands the Sindhi lexicons independent in primary form of morphology which are helpful in recognition of grammatical status of each described lexicon. Stemming words show the root words of Sindhi lexicons. Therefore, hierarchical trees of lexicons may be developed on basis of stemming words. Each tag of UPOS and SPOS is assigned a digital number to process for machine learning model. Sindhi annotated corpus is consisted of six attributes namely, WordID, UPOS, SPOS, WORD, STEM and LEMMA. WordID shows the address of Sindhi

TABLE 4. FEATURE VALUES AND FEATURE NAMES, EXTRACTED FROM SINDHI CORPUS THROUGH TF-IDF

TF-IDF Value	Feature Name
0.141524	سنڌ
0.098352	جو
0.225637	انب
0.306070	گھٹو
0.276330	مٹو
0.080568	آهي
0.213959	گلاب
0.086673	جو
0.223370	گل
0.223370	سھٹو
0.163013	ٿئي
0.113901	ٿو

TABLE 3. DETAIL OF TEXT VECTORS, EXTRACTED FROM SINDHI CORPUS

	Word	Stem	Lemma
Count	3743	3743	3743
Unique	1253	958	1103
Top	جي	جي	جي
Frequency	155	156	155

lexicon, UPOS and SPOS show the universal and Sindhi POS tag sets, Stem shows the stemming words of related Sindhi lexicons and Lemma shows the lemmas of Sindhi lexicons. The corpus is normalized, therefore, there is no missing values available in the annotated corpus. Table 5 shows the records of data set.

Fig. 4 shows the annotation process of UPOS and SPOS, Lemma and Stemming performed on Sindhi corpus document بين کي خوش رکڻ سان ماڻهو پاڻ به خوش رهي ٿو (Bbyan khay khhush rakhann saan maannho paann bi khhush rahay tho). Each Sindhi lexicon is annotated separately with UPOS, SPOS lemma and stemming words.

2.2 Machine Learning Model

To verify and evaluate the annotation of Sindhi corpus for language feature distribution, pattern recognition and lexicon annotation analysis, supervised machine learning model is developed using the SVMs and RF classifiers. SVMs and RFs classifiers are suitable to analyze the text for the purpose of classification tasks and language feature analysis [14-15]. The corpus is partitioned into 80% training dataset and 20% test dataset. The training data set is used to train the model and test data set is used to evaluate the performance of model. Training data set is used to train the model to understand the language features and make familiar it with Sindhi corpus terms

TABLE 5. SINDHI ANNOTATED CORPUS RECORDS

Word ID	UPOS	SPOS	Sindhi Word	Stem	Lemma
2540	1	1	ڪلچر	ڪلچر	ڪلچر
73	6	6	لاءِ	لاءِ	لاءِ
:	:	:	:	:	:
:	:	:	:	:	:
615	1	1	ثابت	ثابت	ثابت
360	4	4	ٿيندو	ٿيند	ٿي

Sindhi Corpus Annotation

Lemma	Stem Suffix	Stem Affix	Stem	SPOS	UPOS	Word	Word ID
بين	بن		بي	ضمير اشارو	DET	بين	1111
کي			کي	حرف جر	ADP	کي	313
خوش			خوش	صفت	ADJ	خوش	333
رک	اڻ		رک	فعل	VERB	رکڻ	1217
سان			سان	حرف جر	ADP	سان	265
مڙيوڻ			مڙيوڻ	اسم	NOUN	مڙيوڻ	252
پاڻ			پاڻ	ضمير	PRON	پاڻ	945
به			به	حرفِ جر	ADP	به	309
خوش			خوش	صفت	ADJ	خوش	333
رهي	اي		ره	فعل	VERB	رهي	330
ٿر			ٿر	فعل	VERB	ٿر	14

FIG. 4. ANNOTATION OF SINDHI TEXT DOCUMENT

therefore, model can easily identify the terms when test set is processed. The model is evaluated and validated with cross validation technique using k-folds, where k = 10. Cross validation technique splits the Sindhi annotated corpus into 10 subsets to analyze and validate the corpus properly. Each round of cross validation analyzes the partitioned part of Sindhi corpus called training data set to assess the annotation process and then validate that analysis on other partitioned part of Sindhi corpus called test data set. This process runs till 10 times on randomly partitioned Sindhi data set to analyze and validate each randomized partition properly. Finally, cross validation process counts the error rate.

RF classifier trains multiple decision trees with target classes and features extracted from Sindhi annotated corpus for classification and finally, it ensembles all the trained trees to give the results. Results are acquired through labelled features UPOS and SPOS. RF is featured with n-estimators=10, criterion = gini.

SVM non-linear classifier classifies the Sindhi corpus on basis of hyperplanes which are decision boundaries. Hyperplanes separate the class elements through boundaries, therefore, SVM non-linear generates multi-hyperplanes to identify the Sindhi tagged lexicons. The proper classification of target classes UPOS and SPOS is done through multi-hyperplanes which describe the accurate results of model and improper classification shows the error rate of model. The features of SVM non-linear are set as: kernel = 'rbf', C=40, gamma = 1. RBF (Radial Basis Function) is one of the kernels of SVM which is suitable for SVM non-linear. The feature gamma is kernel coefficient for 'rbf' which fits the generalization error and over-fitting problem. C is penalty parameter which is used to control the tradeoff between smooth decision boundary and classifying the training points correctly.

3. RESULTS EVALUATION AND ANALYSIS

Results are shown on basis of confusion matrices, accuracy rates, precision, recalls and f-scores. All the measurement techniques are important and significant for the evaluation and analysis of performance of supervised model which performs machine learning operations on Sindhi annotated corpus.

3.1 Confusion Matrix Analysis

There are two classes: UPOS and SPOS; therefore, the performance of both classifiers is different from each other. Fig. 5 shows the confusion matrix of SVM non-linear and UPOS shows the confusion matrix of RF targeting class UPOS. The confusion matrix in Fig. 6 describes the true positive data of Nouns, ADPs, Determiners, NUMs, and PARTs in actual and predicted data. At the same time, it shows false positive data values along with true data of remaining POS items in actual and predicted columns and rows. The confusion matrix of RF shows a better performance in comparison of SVM non-linear. The visual layout of the matrix shows true positive data values of each POS item. There are no false positive values found in the confusion matrix of RF classifier. The evaluation and analysis of confusion matrix shows the better performance of random forest classifier on Sindhi annotated corpus.

SPOS is another class of Sindhi corpus data set. This class represents the Sindhi part of speech tag set.. The performance of machine learning methods is a little bit different from each other targeting class SPOS. The confusion matrix which is derived through SVM non-linear is presented in Fig. 7 and the matrix, derived through RF is shown in Fig. 8.

Visual confusion matrix of SVM non-linear shows true positive values of Nouns, ADPs, Determiners, Adjective, NUMs and presents false positive values of remaining SPOS items. The SPOS items are shown with digital numbers in this matrix. Less number of false positive values in presenting confusion matrix of SVM non-linear, shows a better performance in comparison of same classifier applied on UPOS class. But it does not work better than the RF classifier even targeting SPOS class.

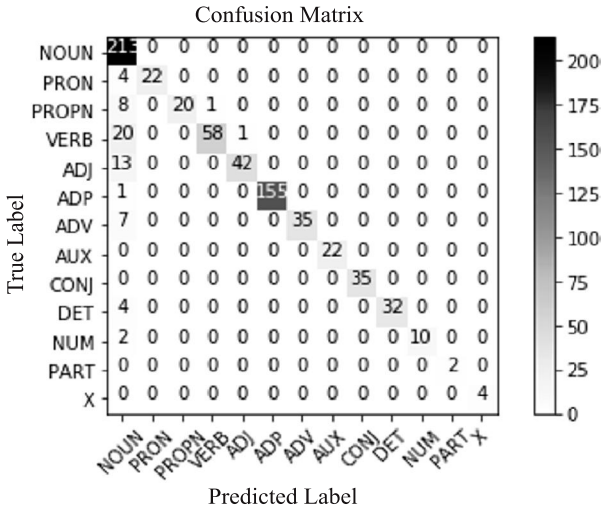

FIG. 5. CONFUSION MATRIX , DERIVED THROUGH SVM NON-LINEAR ALGORITHM TARGETING CLASS UPOS

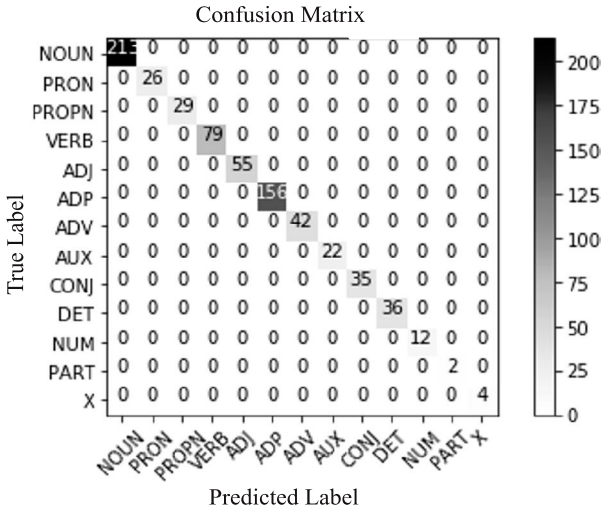

FIG. 6. CONFUSION MATRIX, DERIVED THROUGH RANDOM FOREST ALGORITHM TARGETING CLASS UPOS

Visual confusion matrix of RF classifier (Fig. 8) shows a better performance than the performance of SVM non-linear classifier (Fig. 7) targeting SPOS class. The matrix shows the highest number of true positive data values in actual data and predicted label data. This matrix verifies the better performance and significance of RF classifier applied on Sindhi corpus.

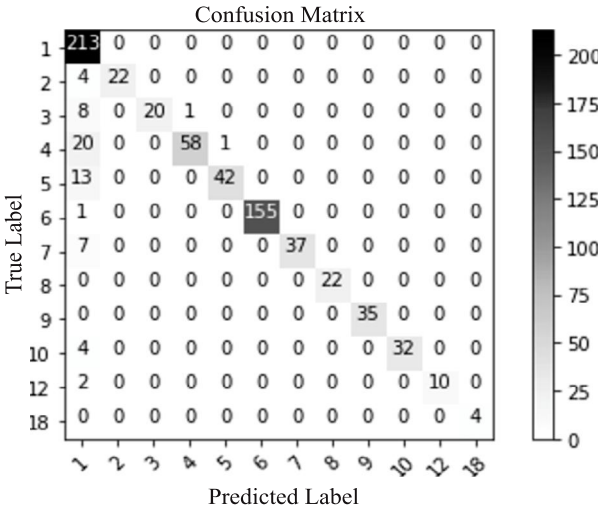

FIG. 7. CONFUSION MATRIX OF SVM NON-LINEAR ALGORITHM TARGETING CLASS SPOS

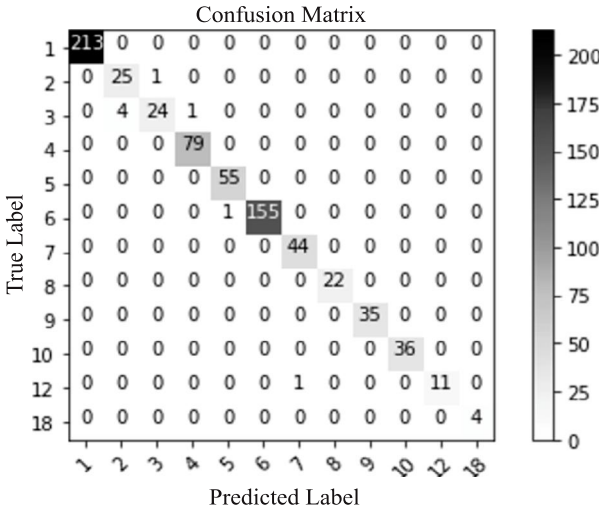

FIG. 8. CONFUSION MATRIX OF RANDOM FOREST ALGORITHM TARGETING CLASS SPOS

3.2 Accuracy Analysis of Supervised Model

The developed model and confusion matrices are evaluated and assessed with accuracy score. Accuracy shows the correctness and incorrectness of working of supervised model. Therefore, accuracy is a mathematical fraction of the accurate classified data and it approves the effectiveness and actual position of the model. Accuracy of the model is acquired through Equation (3).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{\text{TP} + \text{TN} + \text{FN} + \text{FN}} \quad (3)$$

The predictions and decisions are executed on the basis of acquired accuracy of model in this study. The accuracy verifies the performance of confusion matrices derived through the supervised model. Accuracy observes the true values which are classified properly through this study and records the error rate of unclassified data. Therefore, high accuracy of the machine learning methods shows the better efficiency, performance and reliability of model which recognizes the data set for future prediction, pattern recognition and text analysis.

The corpus features are extracted and labelled as class features in the Sindhi corpus. The SPOS gives better results than the UPOS class. The results of accurate classified data are shown in Table 6 with their accuracy rate.

Accuracy of classification methods shows the proper selection of terms and true annotation of corpus. Supervised model shows different results of both machine learning classifiers on Sindhi annotated corpus. SVM non-linear machine learning method shows less accuracy in comparison of RF machine learning method in all classes. Thus, results show proper annotation of Sindhi part of speech to Sindhi corpus. There is a big difference of performance between both classifiers on non-English linguistic corpus. This corpus is using Unicode-8 code for Sindhi lexicons. Fig. 9 shows the difference of performance of both machine learning methods applied on multi-class based Sindhi corpus.

TABLE 6. PERFORMANCE OF MACHINE LEARNING METHODS ON SINDHI ANNOTATED CORPUS USING MULTI-CLASSES

Method	Class UPOS Accuracy (%)	Class SPOS Accuracy (%)
SVM Non-Linear	89.16	89.1
Random Forest	99.57	99.89

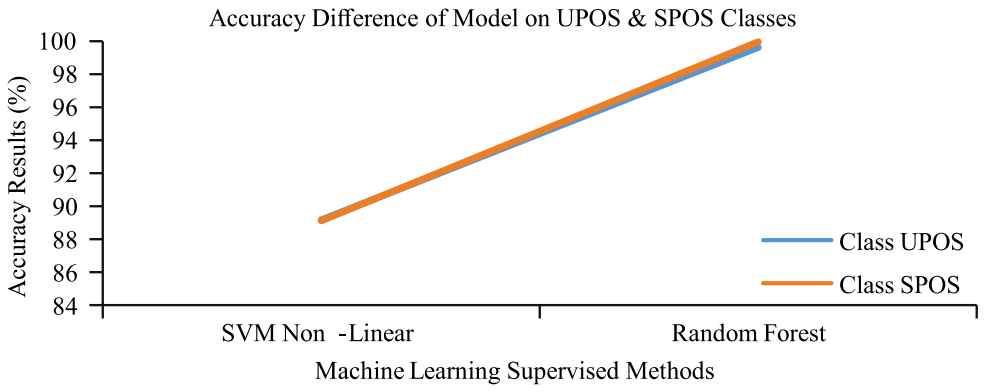

FIG. 9. COMPARISON OF ACCURACY OF MACHINE LEARNING METHODS

3.3 Precision, Recall and F-Measure Analysis

The precision and recall show the percentage of relevant and irrelevant data available in Sindhi annotated corpus, they confirm the true predicted data values. Generally, precision ratio is acquired from the number of related or relevant data instances which are obtained from total number of relevant and irrelevant data, whereas recall shows the sensitivity ratio of data by acquiring relevant data from total number of relevant data. The F-score is a measurement of test accuracy that is weight harmonic mean of precision and recall. Table 7 shows precision, recall and F-score of Sindhi corpus labelling class UPOS and Table 8 shows precision, recall and F-score of Sindhi corpus labelling class SPOS.

Precision, recall and F1 score of UPOS and SPOS acquired by RF are better than the SVM non-linear. The average precision, recall and F-score results of class UPOS are shown in Table 9, which describe the better performance of RF supervised method on Sindhi annotated corpus.

Average precision, recall and F-score results of class SPOS are shown in Table 10, which describe the better performance of RF supervised method in comparison of SVM non-linear.

Fig. 10 shows the dissimilarity of results of Precision, Recall and F-score results. Results of random forest show the high number of relevant data instances with annotated labelled classes. It shows the proper distribution of Sindhi lexicons in corpus.

TABLE 7. COMPARISON OF PRECISION, RECALL AND F-SCORE TARGETING CLASS UPOS

Predicted Class Elements	Precision (%)		Recall (%)		F-Score (%)	
	SVM Non-Linear	Random Forest	SVM Non-Linear	Random Forest	SVM Non-Linear	Random Forest
NOUN	0.78	100	100	100	0.88	100
PRON	100	100	0.85	100	0.92	100
PROPN	100	100	0.69	100	0.82	100
VERB	0.98	100	0.73	100	0.84	100
ADJ	0.98	100	0.76	100	0.86	100
ADP	100	100	0.99	100	100	100
ADV	100	100	0.83	100	0.91	100
AUX	100	100	100	100	100	100
CONJ	100	100	100	100	100	100
DET	100	100	0.89	100	0.94	100
NUM	100	100	0.83	100	0.91	100
PART	100	100	100	100	100	100
UNKNOWN	100	100	100	100	100	100

TABLE 8. COMPARISON OF PRECISION, RECALL AND F1 SCORE TARGETING CLASS SPOS

Predicted Class Elements	English Equivalence	Precision (%)		Recall (%)		F-Score (%)	
		SVM Non-Linear	Random Forest	SVM Non-Linear	Random Forest	SVM Non-Linear	Random Forest
اسم (Ism)	NOUN	0.78	100	100	100	0.88	100
ضمير (Zameer)	PRON	100	0.96	0.85	100	0.92	0.98
اسم خاص (Ism khaas)	PROPN	100	100	0.69	0.97	0.82	0.98
فعل (Fael)	VERB	0.98	100	0.73	100	0.84	100
صفت (Sifat)	ADJ	0.98	100	0.76	100	0.86	100
حرف جر (Harf e jar)	ADP	100	100	0.99	100	100	100
ظرف (Zarf)	ADV	100	100	0.84	100	0.91	100
فعل معاون (Fael Maawan)	AUX	100	100	100	100	100	100
حرف جملو (Harf jumlo)	CONJ	100	100	100	100	100	100
ضمير اشارو (Zameer ishaaro)	DET	100	100	0.89	100	0.94	100
صفت عددي (Sift adadi)	NUM	100	100	0.83	100	0.91	100
نامعلوم (Naa maaloom)	UNKNOWN	100	100	100	100	100	100

TABLE 9. PRECISION, RECALL AND F1 AVERAGE SCORE TARGETING CLASS UPOS

Method	Precision (AVG)	Recall (AVG)	F1-Score (AVG)
SVM Non-Linear	93	91	91
Random Forest	100	100	100

TABLE 10. PRECISION, RECALL AND F1 AVERAGE SCORE TARGETING CLASS SPOS

Method	Precision (AVG)	Recall (AVG)	F1-Score (AVG)
SVM Non-Linear	93	91	91
Random Forest	100	100	100

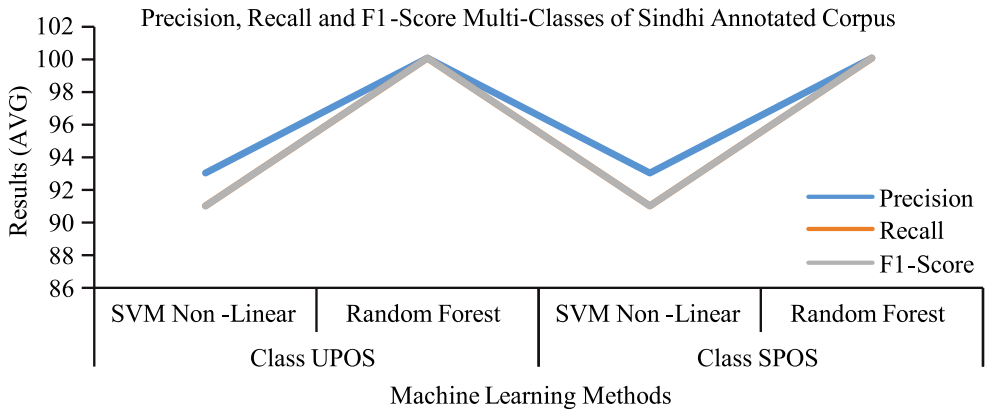

FIG. 10. DISSIMILARITY OF RESULTS OF PRECISION, RECALL AND F-SCORE

4. CONCLUSION

The presented research study has developed a novel Sindhi annotated corpus for the purpose of more research on Sindhi corpus. Study has performed supervised classification on Sindhi annotate corpus to assess the accuracy of traditional machine learning approaches to solve the NLP problems of Sindhi language. Supervised regarding machine learning methods are evaluated and assessed with 10-fold cross validation. The Sindhi annotated corpus is segmented into 80% training dataset and 20% test dataset. The machine is trained with 80% training dataset. Each fold of cross validation has processed to partition corpus into subsets to analyze the training set and validate the test set. All processes of cross validation have done randomly. The study observes the performance of RF machine learning method better than the SVM non-linear on basis of obtained results. The performance is evaluated through confusion matrices, accuracy, precision, recall and F1-score. The high accuracy rate of machine learning methods shows the proper annotation of Sindhi corpus, thus corpus may be utilized for other NLP researches and purposes.

Sindhi NLP resources may also be used for further research on Sindhi language feature identifications and variations, topic modeling, sentiment and semantic analysis and information retrieval.

ACKNOWLEDGEMENT

This research study is part of doctoral research work on "Sentiment Analysis for Sindhi Text", continue at SZABIST (Shaheed Zulfiqar Ali Bhutto Institute of Science & Technology), Karachi, Pakistan. Authors acknowledge the support of Dr. Hussnain Mansoor Ali Khan, Program Coordinator and Dr. Imran Amin, Department of Computer Science, for provision of resources and facilities at SZABIST Karachi Sindh Pakistan.

REFERENCES

[1] Kesavaraj, G., and Sukumaran, S., "A Study on Classiûcation Techniques in Data Mining", Proceedings of IEEE 4th International Conference on Computing, Communications and Networking Technologies, pp. 1-7, 2013.

[2] Abro, M.A., Nawaz, D.N., and Abro, W.A., "Performance Analysis of Dissimilar Classification Methods Using RapidMiner", Sindh University Research Journal (Science Series), Volume 48, No. 1, pp. 185-188, Jamshoro, Pakistan, 2016.

[3] Motlani, P., "Developing Language Technology Tools and Resources for a Resource-Poor Language: Sindhi", Proceedings of NAACL: NAACL-HLT, pp. 51-58, 2016.

[4] Mahar, J.A., and Memon, G.Q., "Rule Based Part of Speech Tagging of Sindhi Language", IEEE International Conference on Signal Acquisition and Processing, pp. 101-106, 2010.

[5] Mahar, J.A., Shaikh, H., and Memon, G.Q., "A Model for Sindhi Text Segmentation into Word Tokens", Sindh University Research Journal (Science Series), Volume 44, No. 1, pp. 43-47, Jamshoro, Pakistan, 2012.

[6] Mahar, J.A., and Memon, G.Q., "Sindhi Part of Speech Tagging System using WordNet", International Journal of Computer Theory and Engineering, Volume 2, No. 4, pp. 538, 2010.

[7] Dootio, M.A., and Wagan, A.I., "Syntactic Parsing and Supervised Analysis of Sindhi Text", Journal of King Saud University – Computer and Information Sciences, [DOI:10.1016/j.jksuci.2017.10.004], 2017.

[8] Motlani, R., Lalwani, H., Shrivastava, M., and Sharma, D.M., "Developing Part-of-Speech Tagger for a Resource Poor Language: Sindhi", Proceedings of 7th Conference on Language and Technology, Poznan, Poland, 2015.

[9] Motlani, R., Tyers, F.M., and Sharma, D.M., "A Finite-State Morphological Analyzer for Sindhi", Proceedings of 10th International Conference on Language Resources and Evaluation, 2016.

[10] Siraj, "Sindhi Boli", 2nd Edition, Sindhi Language Authority, Hyderabad, Sindh, Pakistan, 2009.

[11] Bag, M.K., "Sindhi Vyakaran", Sindhi Adabi Board, Jamshoro, Sindh, Pakistan, 2015.

[12] Petrov, S., Das, D., and McDonald, R., "A Universal Part-of Speech Tag Set", arXiv Preprint arXiv:1104.2086, 2011.

[13] Taylor, A., Mitchell, M., and Beatrice, S., "The Penn Treebank: An Overview", Treebanks, pp. 5-22. Springer, Dordrecht, 2003.

[14] Sarker, A., and Graciela, G., "Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training", Journal of Biomedical Informatics, Volume 53, pp. 196-207, 2015.

[15] Onan, A., Serdar, K., and Hasan, B., "Ensemble of Keyword Extraction Methods and Classifiers in Text Classification", Expert Systems with Applications, Volume 57, pp. 232-247, 2016.