POS-HOML: POS Tagging Technique For Gujarati Language Using Hybrid Optimal And Machine Learning Approaches

Pooja M Bhatt[1], Dr. Amit Ganatra[2]

[1]Research scholar, Computer Engineering department, Charusat University, Changa, India
[2]Dean, faculty of engineering Computer Engineering department, Charusat University, Changa, India.

[1]bhattpooja.393@gmail.com, [2] amitganatra.ce@charusat.ac.in

Abstract — *Natural language processing facilitates the interaction between humans and machines. The primary use of the POS is to recognize words' tags, such as nouns, verbs, and adjectives. For the Indian language, it is a difficult task to allocate the correct POS tag to each word in a judgment because of some unknown words in Indian languages. The earlier work for Indian languages was dependent on statistical and rule-based approaches. The Statistical approaches used mathematical equations, while the rule-based approach needs precise language knowledge and hand-written rule. This paper suggests the POS category method for Gujarati language using hybrid optimal and machine learning techniques (POS-HOML) to improve POS tagging. The first contribution of the proposed POS-HOML is to introduce optimal feature selection, which optimizes the multiple features to avoid dimensionality problems. The second contribution is applying the various machine learning techniques, like hidden Markov model (HMM), rule-based approach, Hybrid (combination of rule and Hidden Markov model), Recurrent neural network (RNN), Conditional random field (CRF), Long Short-Term Memory (LSTM) to classify the POS of the given text. Finally, the paper compares various methods using standard bench datasets to analyze the effectiveness of other POS methods in terms of accuracy, precession, recall, F-measure.*

Keywords — *POS tagging, Gujarati language, optimal, machine learning, hidden Markov model, rule-based network, Long Short-Term Memory, deep neural network*

I. INTRODUCTION

India is an excellent multilingual country with different cultures. It has many languages in paper form and more than a thousand spoken languages. The foundations of India recognize 22 languages spoken in different parts of the fatherland. Indo-Aryan and Dravidian languages can divide into two primary language families. There are some essential differences between these language classes [1]. Their ways of mounting words and syntax are dissimilar. Humans have created complex social, political, and technological systems through decision-making and rational and irrational processes. Natural language certainly contributed to the effective presentation of specific human tasks. POS mark delivery is considered one of the main tools in many natural language dispensation programs such as word suppression, sequence retrieval, information processing, analysis, survey, and machine translation [2][3].

POS tagger represents part of the conversation and the other markers in the class for each word in the corpus.[5] [6]. One of the main functions of a natural language dispensation program, such as speech gratitude, is in order retrieval, machine translation, grammar testing, and word comprehension. This helpful information varies depending on the specific NLP program (data recovery, machine translation) [7]. Tagging is a source of many research challenges. The type and level of these challenges depend on the language under thought. Several methods are helpful for POS classification. Most modern methods depend on machine learning [9] [11]. In a rule-based approach, hand-written rules are helpful to distinguish between ambiguities of identity [12].
Product Probability and Significance Sequence Probability are synchronous signs HMM based on tag sequence selection that combines target-based product trees or probability characteristics with maximum entropy patterns [13]. Some of the reasons for the relative lack of rule-based approaches are machine readings and reduced hardware capacity (processor, memory, disk space), and researchers recommend building-based zip codes. The performance of the POS tagging model depends on corpus data, which is the majority, deliberate, and frequently used mark method trained in HMM [14]. Some modern signs combine two or more approaches and are designed to increase overall accuracy. For example, people in Gujarat frequently use Gujarati for communication. The POS label plays an essential role in developing natural languages processing programs such as analyzer and morphological analyzer. The literature contains numerous articles on POS industrial languages such as Hindi, Marathi, Oriya, and Punjabi. However, the content of the Gujarati text is problematic. Adaptive approaches are helpful for rich languages with exclusive conversions [15].
There are two main types of coding: supervision and control. In the first case, the process applied the manual interpretation, and then the process used algorithms that state that 95% or more accuracy ratios are not uncommon for the POS indicator that marks the hull. In the second

case, set a mark with induction that does not indicate the corpus for their training data [16]. The problem with this approach is that it can give signs of sentences that are not adequate according to the grammatical rules of the language. A hybridization is a different approach. The Hybrid may work better than a statistical or rule-based approach. The hybrid advance first uses the probability features of the arithmetical system and the manually oblique linguistic rules [17]. Therefore, the primary purpose of this study is to facilitate the development of NLP research in a language that incorporates standard and rule-based approaches to test whether further accuracy of the POS mark can achieve with a hybrid approach. To do this, first, run a synchronous POS tag based on the hidden Markov model (HMM), and then combine the transformation rules into two known words to see how the accuracy of the POS tag changes [18] [19].

II. RELATED WORK

Pecheuxet al. [21] reviewed two existing learning proposals with vague etiquette, which students generally control weakly. Meanwhile, the condition is a continuation of the random field model. Focus on some voice tags, but given the large number of 10 languages, uncertainty (a) is guaranteed if both single and bilingual resources are available. It indicates that good performance can be achieved even in this case. (b) Two students use different training traits and succeed in different situations. (c) Aside from choosing the correct learning method, many other factors can improve performance in critical language exchange environments.

Alhasan et al. [22] proposed an efficient classification move toward using the Bee Colony Optimization algorithm for the Arabic language. The difficulty is presented as a map, indicating new ways to get points with possible signs of the sentence, and the bees find the most excellent way. The planned move toward is estimated using the 18 million words KALIMAT Corpus. The inexperienced outcome shows that the proposed Hybrid, hidden Markov model and rule-based methods achieved 98.2% accuracy with 98%, 97.4%, and 94.6%, respectively. In addition, the planned move toward identifying all the signs presented to the building, while the mentioned approach can recognize only three signs.

Khan et al. [23] proposed a novel category advance using linear-chain conditional random fields (CRF). Their work is the first example of the CRF approach to representing Urdu POS. The planned model uses a powerful, consistent and consistent language-independent feature and a set of linguistic features. Linguistic features include the audio part of the preceding word and the suffix of the current word, and language-independent features include the "context word window." The move toward values for the Urdu POS sub-vector technology, which is considered a sophisticated two-level database. The results show a CRF move toward improving the F level by 8.3–8.5% in previous trials.

Schulz et al. [24] investigate to explore resource-poor, diverse, non-standard language coping strategies in the field of ordinary speech dispensation. Adequate annotation resources can cause problems in deck training, and additional training data are available to modify existing resources effectively. The resulting POS sample achieves approximately 91% accuracy in different test modules representing different types, terms, and MHG types. Magistry et al. [25] presented experiments in part-of-speech tagging of low-resource languages. Sometimes data named in the target language and the bilingual corpora are not available. They want the target language to be close to the language with the best sources. They are testing the French language in three regions.

Myint et al. [26] studied to clarify the POS of the word given in Myanmar texts. The goal is to eliminate this ambiguity in Burmese and assign a separate POS to apiece word in the verdict. The subsequent idea illustrates this. (I) Create sentences and divide them into words using the rules of writing and the approach that best applies to the Myanmar monolingual dictionary. (Ii) Use Joint Entropy (JE) for sale. The importance of simultaneous probabilities can attribute to the free order and structure of words for an effective and accurate definition of POS in Myanmar text. Six hundred twenty phrases and 15,000 words were formed using Myanmar's marked text in a single language and a selected dictionary.

Mohammed et al. [27] presented various machine learning approaches (HMM and CRF) and neural network models to display statistical POS taggers in Somalia. Somalia's POS taggers surpass 87.51% of the latest status POS tags based on ten cross-checks. The main contributions of this study are (1) the creation of common point-of-sale tags, (2) the comparison of performance with existing state-of-the-art technology, and (3) the study of word insertion used in Somali POS taggers. De Oliveira et al. [28] studied state-of-the-art POS-tagging surroundings for Brazilian Portuguese clinical texts. They reviewed several neural network-based POS tagging algorithms, but there was no special algorithm for Portuguese medical text, so the Flare tool was helpful for the exceptional performance of the press domain. They have done the normalization process for several domain companies (a new company that includes two journals, one biomedicine, one medical, and three). Flair algorithm was prepared for all domains, and five domains were experiments for all domains. The clinical model achieves 92.39% accuracy (preceding clinical POS category clinical work achieved 91.5%), and the biomedical sample achieves 97.9% accuracy. Their test instrument evaluated all models.

Akhil et al. [29] proposed an in-depth learning-based approach to coding parts of Malay dialogue. Experiments on authentic databases show that the proposed method is more accurate and precise than some existing methods. This process is one of the pre-processing stages of many natural language processing tasks. Early approaches depended on simple horoscopes, but several methods were published in the literature, including machine learning techniques such as artificial neural networks. Feng et al. [28] proposed neural machine translation (NMD) performance using syntax in target languages (such as

POS), which is more accurate and delayed than in-depth syntax such as pro-analysis.

Besharati et al. [31] presented word vectors used to make POS reference in Persian for MLP and LSDM neural networks and contrast the neural model's consequences with the actual benchmark HMM. They also used a bilateral LSDM neural network to study the effect of bilateral learning mechanisms on Persian POS identities. The outcomes of the various samples in this study show that neurological models effectively predict the correct POS tags for OOV terms. Therefore, they proposed a hybrid model representing HMM and a single-layer bilateral LSDM model representing pioneering POS. This hybrid model has been triumphant in upgrading both the HMM and the neural models with 97.29% accuracy.

Maulana et al. [32] presented the domain adaptation method with an additional dictionary built into the link rules. A specific domain is helpful for the domain of beauty products. One component of this system is the POS tagger with unnamed vocabulary from common domains and target domains. Word classes are helpful in the target domain dictionary based on the balance between link information and manual tag. The dictionary was urbanized in Indonesian and English based on the database observation because the words were primarily in English. The enabled glossary is further to the dictionary from the original POS Tagger, and POS Tagger uses the public domain to provide domain-specific information. Tags with additional dictionaries attain 68.99% correctness, and the proportion of words documented by tags is 92.36%.

III. PROBLEM METHODOLOGY AND SYSTEM ARCHITECTURE

A. Problem methodology

The Gujarati language is one of the most spoken languages in Gujarat, and POS tagging plays a pivotal role in developing NLP applications like Parser and Morphological analyzer. However, the basic models of translating neural machines depend on a constant increase in the size of the parameters to accomplish optimal presentation that is not useful for a mobile phone. Few articles are available on POS tagging tasks for Indian languages like Hindi, Marathi, Odisa, Panjabi. The syntactic categories assigned to words in a judgment are called POS tagging dilemmas, which play an essential role in many NLP and sequence repossession functions. Recently, machine and deep learning methods have primarily helped to avoid those problems. The researchers have proposed the HMM-based POS tagger for the Gujarati language, requiring extensive domain and linguistic knowledge and resources. Despite the variety of methods, posing is a sign of many challenges that require new solutions. Theoretically, one solution is to turn the tag problem into an optimization problem for well-defined needs and to use evolutionary methods to solve the optimization problem. The POS tagging method experiments for the Gujarati language using hybrid optimal and machine learning techniques (POS-HOML) to address these challenges. The main objective of the proposed POS-HOML method is the list as follows:

- A hybrid technique is an exercise for POS taggers in the Gujarati language.
- To introduce a novel optimization algorithm for optimal feature selection, which minimizes the dimensionality problems.
- To use various techniques like HMM, CRF, RNN, Rule-based, LSTM for Tag classification for Gujarati text input and explore the comparative study analysis of the accuracy of the various tagger.

B. System architecture

It is essential to convert the original text into a format that can be understood and used by machine learning methods called text pre-processing to work with text data. A POS tagging is a particular sign assigned to each word as part of a text and often other grammatical types, such as anxiety, number (plural/singular). Letter-like POS marks are helpful in body search and text analysis tools. At the feature acquisition stage, the researchers need to do a part of the speech to separate names/phrases from the reviews that may be product features. The researchers use the BIS-POS tag to analyze each sentence and mark the spoken part of each word, noun, adjective, verb, adverb. After that, the researchers select particular features using various contexts like word length, Suffix, Prefix. Then the tag is classified into noun, verb, adjective, pronoun, conjunction, using variously supervised, Unsupervised, Hybrid and Deep learning approaches. Fig. 1 shows the system architecture of the proposed POS-HOML method.

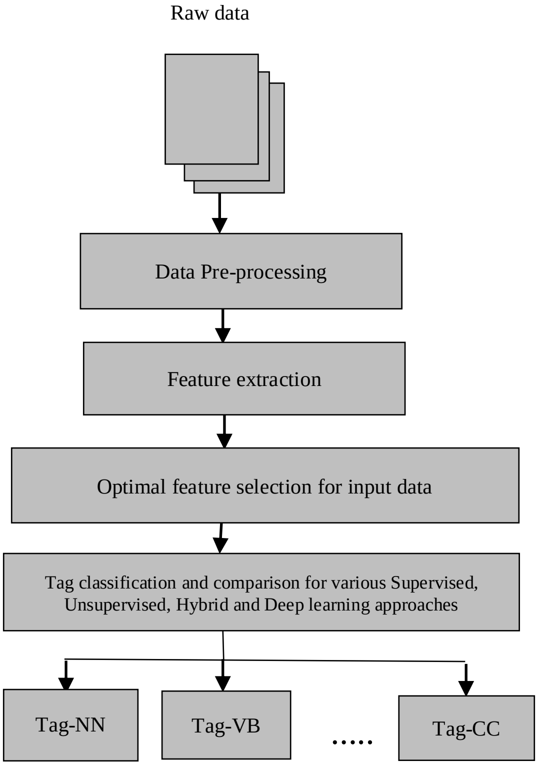
Raw data
Data Pre-processing
Feature extraction
Optimal feature selection for input data
Tag classification and comparison for various Supervised, Unsupervised, Hybrid and Deep learning approaches
Tag-NN
Tag-VB
.....
Tag-CC

Fig. 1System architecture of proposed POS-HOML method

The proposed methodology is divided into two parts:

a) Optimal Feature selection for input data

The choice of functions plays an essential role in POS cataloguing. The main features of the POS code are elected based on a different combination of language and tags. Indian language suffix and prefix properties are helpful features of POS code. The researchers discuss a combination of different features to determine the characteristic for POS classification tasks. The features focused in the input text data are Symbol, Length of the word, frequent word, Suffix, Prefix, and presence of special characters**.** Here the researchers applied optimal feature selection based on the optimal feature selection approach, which optimizes the multiple features to avoid dimensionality problems.

b) POS tagging classification and comparison for various Supervised, Unsupervised, Hybrid and Deep learning approaches

Supervised methods use features from labelled training facts, including learning examples, while unsupervised methods use advanced computational methods. The hybrid approach uses a combination of supervised and unsupervised approaches, while Deep learning uses multiple levels of representation to extract higher-level capabilities from the raw input text data by gradually processing data.

In a supervised approach, HMM calculates the probability of an array of tags. HMM refers to the most appropriate sign for a word or token in a sentence based on the Markov chain property. Rule-based POS captioning is a well-known solution for identifying tags in words using predefined rules. However, many analysts prefer statistically-based approaches as well as rule-based methods for better experimental accuracy. The researchers applied the hidden Markov model with a rule-based approach, a hybrid approach to classifying the POS of the given input sentence.

Conditional Random Field(CRF) calculates tagging probabilities using non-independent features of given input text. RNN is the first algorithm that recollects its input because of an internal reminiscence, making it ideal for the device to gain knowledge of issues involving sequential information. Long Short-Term Memory (LSTM) networks are a form of RNN that can gain knowledge of order dependence in collection prediction troubles.

IV. RESULTS AND EVALUATION

In this segment, the researchers evaluate the presentation of the various classifier method using a standard benchmark dataset. The performance of the Hybrid classifier compared with the other methods are HMM, CRF, rule-based, RNN, and LSTM, in requisites of accuracy, precession, recall, and F-measure.

A. Dataset description

Indian languages corpora initiative (ILCI) launched the Communications and Information Technology (MCIT) for the technology development for Indian languages (TDIL) to establish a parallel corporation of Indian languages, including English as well as Hindi, Bengali, Marathi, Gujarati. The second phase includes 11 other major Indian languages (22 block or national languages in India, English the official language). For this experiment, the researchers use Gujarati corpus from the Multilingual database of TDIL, including various data about entertainment, art and culture, sports, philosophy, religion, science & technology, and sports. Table 1 describes dataset details with the number of words.

The performance of proposed various approaches like HMM, CRF, rule-based, RNN and LSTM are calculated in provisions of accuracy (A), precession (P), recall (R) and F-measure (F). The details of performance metrics are discussed as follows:

$$A = \frac{\#correct\ tagged\ words}{\#total\ words} \quad (1)$$

$$P = \frac{\#correct\ ed\ answer}{\#total\ words} \quad (2)$$

$$R = \frac{\#correct\ ed\ answer\ specified\ by\ system}{\#total\ words} \quad (3)$$

$$F = \frac{R \times P}{R + P} \quad (4)$$

Table 1Dataset description

Sr. No.	Corpus	Number of words
1	Guj_artsand culture	14k
2	Guj_economy_set	23.1k
3	Guj_entertainment_set	207K
4	Guj_philosophy_set	29k
5	Guj_religion_set	31k
6	Guj_science&tech_set	66.8K
7	Guj_sports_set	36K

B. Comparative analysis of various approaches

Figure 2 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_arts and culture_set." Figure 2 depicts the accuracy of the Hybrid classifier as 82% While 97%,91%,82%,62%, and 84% for HMM, CRF, Rule-based, LSTM, and RNN classifier,respectively. The precession of the proposed Hybrid classifier is 83%, while 97%,91%,81%,97%, and 91% HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 2. The recall of the Hybrid classifier is 84%, while 97%, 90%,82%, 74%, and 75% HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 2 also describes the F-measure of the Hybrid is 87% while 97%, 89%, 82%,84%, and 86% HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

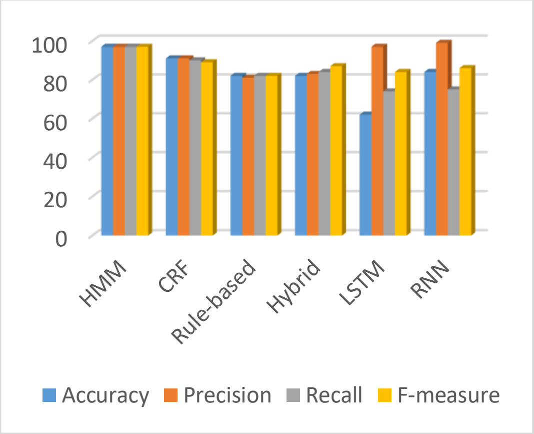

Fig. 2 Comparative analysis summary of various machine and deep learning approaches for "Guj_arts and culture_set."

Figure 3 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_economy_set." Figure 3 depicts the accuracy of the Hybrid classifier is 72%, While 73%,72%,71%,76%, and 75% for HMM, CRF, Rule-based, LSTM, and RNN classifiers, respectively. The proposed hybrid classifier precession is 69%, while 91%,73%,72%,67%, and 65% HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 3. The recall of the Hybrid classifier is 73 while 73%, 78%,70%, 95%, and 95% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 3 also describes the F-measure of the Hybrid is 72% while 77%, 73%, 71%,78%, and 77% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

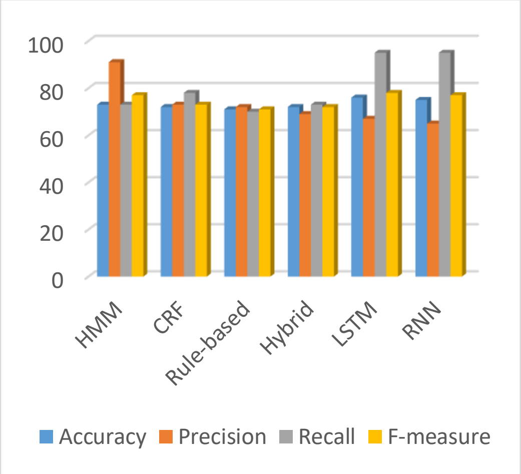

Fig. 3 Comparative analysis summary of various machine and deep learning approaches for "Guj_economy_set"

Figure 4 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_entertainment_set." Figure 4 depicts the accuracy of the Hybrid classifier is 82%, While 77%,91%,80%,90%, 87% for HMM, CRF, Rule-based, LSTM, and RNN classifier respectively. The proposed hybrid classifier precession is 83%, while 87%,90%,81%,99%, and 92% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 4. The recall of the Hybrid classifier is 80%, while 77%, 92%,78%, 86%, and 87% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 4 also describes the F-measure of the Hybrid is 79% while 79%, 90%, 80%,92%, and 89% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

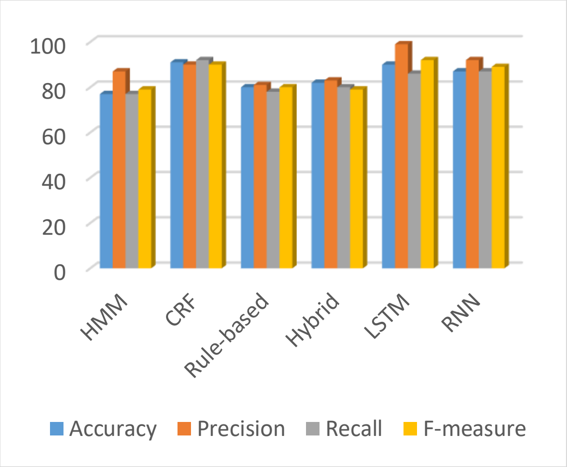

Fig. 4 Comparative analysis summary of various machine and deep learning approaches for "Guj_entertainment_set"

Figure 5 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_philosophy_set." The Figure 5 depicts the accuracy of the Hybrid classifier 71% While 63%,70%,68%,72% and 75% forHMM, CRF, Rule-based, LSTM, and RNN classifier respectively. The precession of the proposed Hybrid classifier is 72%, while 90%,68%,67%,94%, and 98% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 5. The recall of the Hybrid classifier is 69%, while 63%, 70%,63%, 63%, and 63% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 5 also describes the F-measure of the Hybrid is 72% while 67%, 69%, 69%,75%, and 77% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

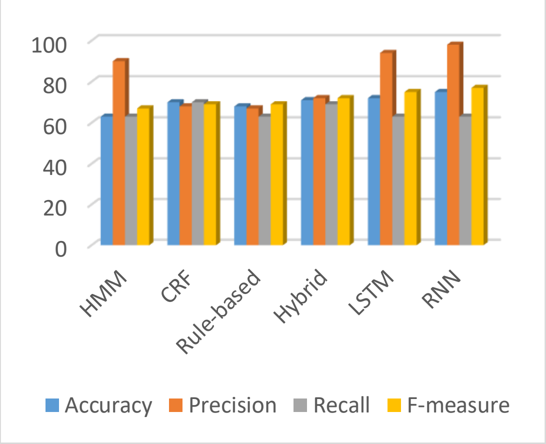

Fig. 5 Comparative analysis summary of various machine and deep learning approaches for "Guj_philosophy_set"

Figure 6 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_religion_set." The Figure 6 depicts the accuracy of the Hybrid classifier is 70% While 62%,67%,69%,74%,73% for HMM, CRF, Rule-based, LSTM, and RNN classifier respectively. The proposed hybrid classifier precession is 74%, while 87%,63%,72%,94%, and 90% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 6. The recall of the Hybrid classifier is 75%, while 62%, 67%, 66%, 66%, and 71% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 6 also describes the F-measure of the Hybrid is 71% while 65%, 68%, 68%,77%, and 79% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

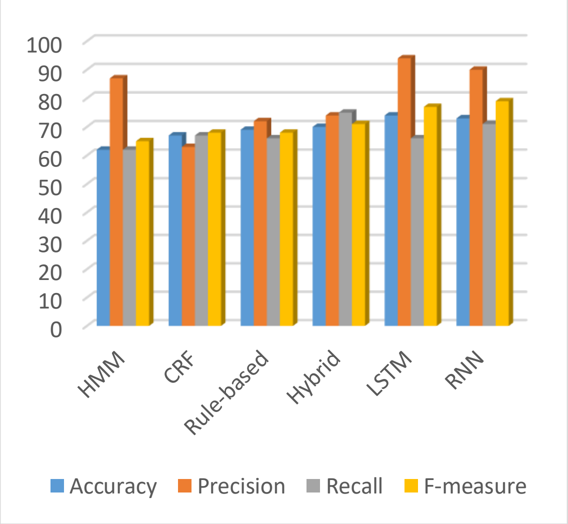

Fig. 6 Comparative analysis summary of various machine and deep learning approaches for "Guj_religion_set"

Figure 7 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_science and tech_set." Figure 7 depicts the accuracy of the Hybrid classifier is 80%, While 78%,79%,76%,82%and 89% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. The proposed hybrid classifier precession is 78%, while 92%,91%,78%,98%, and 98% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 7. The recall of the Hybrid classifier is 76%, while 78%, 75%,76%, 71%, and 80% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 7 also describes the F-measure of the Hybrid is 80% while 81%, 82%, 77%,82%, and 88% HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

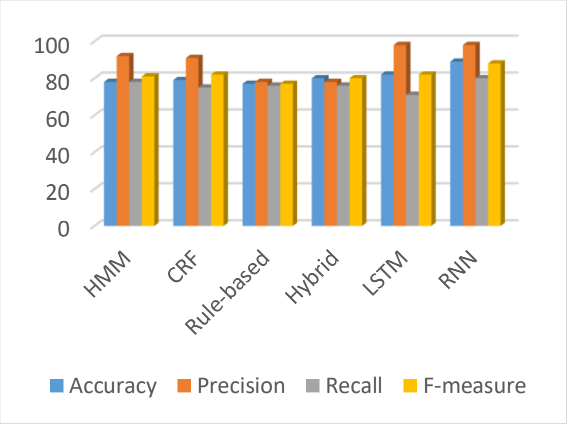

Fig. 7 Comparative analysis summary of various machine and deep learning approaches for "Guj_science and tech_set"

Figure 8 describes the performance compassion of the various approaches like HMM, CRF, rule-based, Hybrid, RNN, and LSTM classifier for "Guj_sports_set." Figure 8 depicts the accuracy of the Hybrid classifier is 73%, While 71%,72%,70%,80%, and 81% for HMM, CRF, Rule-based, LSTM, and RNN classifier respectively. The proposed hybrid classifier precession is 89%, while 90%,89%,73%,98%, and 99% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively, as shown in Figure 8. The recall of the Hybrid classifier is 75%, while 74%, 75%,73%, 70%, and 72% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively. Figure 8 also describes the F-measure of the Hybrid is 76% while 78%, 76%, 70%,82%, and 83% for HMM, CRF, Rule-based, LSTM, and RNN classifier, respectively.

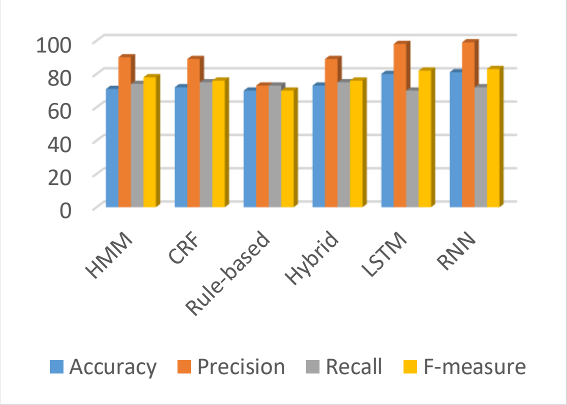

Fig. 8 Comparative analysis summary of various machine and deep learning approaches for "Guj_sports_set"

V. CONCLUSIONS

The researchers have proposed a POS tagging method for the Gujarati language using optimal feature selection and explored comparative analysis of various tagging techniques. The tagging approach depends on various factors like dataset length, suffix, prefix, length of words. The simulation results show that the average accuracy of the Hybrid classifier is2% higher than the HMM and Rule-based classifiers, While 5% lower than the RNN classifier and LSTM classifier. The average precision of the Hybrid classifier is 4% higher than the HMM and Rule-based classifiers while 12% lower than the LSTM classifier and RNN classifier. The average recall of Hybrid classifiers is 3% higher than HMM and Rule-based classifiers, 1% lower than LSTM classifier and RNN classifier. The average F-measure of the Hybrid classifier is 3% higher than the HMM and Rule-based classifiers while 5% lower than the LSTM classifier and RNN classifier. Thus, the Hybrid model works better than the HMM and Rule-based approach, while optimal features selection help CRF for accuracy improvement. Deep learning can automatically optimize the feature by using neural network architecture, which is helpful to improvise the POS taggers better than the Hybrid approach.

REFERENCES

[1] Krishnapriya, V., P. Sreesha, T. R. Harithalakshmi, T. C. Archana, and Jayasree N. Vettath. Design of a POS tagger using conditional random fields for Malayalam. In 2014 First International Conference on Computational Systems and Communications (ICCSC), (2014) 370-373. IEEE.

[2] Forsati, R. and Shamsfard, M., 2014. Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches. Applied Mathematical Modelling, 38(13) 3193-3211.

[3] Nongmeikapam, K. and Bandyopadhyay, S., 2012. A transliteration of CRF based Manipuri pos tagging. Procedia Technology, 6 (2012) 582-589.

[4] Crespo, M. and Frías, A., Stylistic authorship comparison and attribution of Spanish news forum messages based on the TreeTagger POS tagger. Procedia-Social and Behavioral Sciences, 212 (2015) 198-204.

[5] Alex, M. and Zakaria, L.Q., Kadazan part of speech tagging using transformation-based approach. *Procedia Technology*, *11* (2013) 621-627.

[6] Antony, P.J., Mohan, S.P. and Soman, K.P., SVM based part of speech tagger for Malayalam. In 2010 International Conference on Recent Trends in Information, Telecommunication and Computing (2010) 339-341, IEEE.

[7] Bach, N.X., Hiraishi, K., Le Minh, N. and Shimazu, A., Dual decomposition for Vietnamese part-of-speech tagging. Procedia Computer Science, 22 (2013) 123-131.

[8] Brett, D. and Pinna, A., 2015. Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism. Procedia-Social and Behavioral Sciences, 198, 52-57.

[9] Carneiro, H.C., França, F.M. and Lima, P.M., Multilingual part-of-speech tagging with weightless neural networks. Neural Networks, 66 (2015) 11-21.

[10] Liu, K., Chapman, W., Hwa, R. and Crowley, R.S., Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. Journal of the American Medical Informatics Association, 14(5) (2007) 641-650.

[11] Losee, R.M., Natural language processing in support of decision-making: phrases and part-of-speech tagging. Information processing & management, 37(6) (2001) 769-787.

[12] Sánchez-Martínez, F., Pérez-Ortiz, J.A. and Forcada, M.L., Using target-language information to train part-of-speech taggers for machine translation. Machine Translation, 22(1) (2008) 29-66.

[13] Han, C.H. and Palmer, M., 2004. A morphological tagger for Korean: Statistical tagging combined with corpus-based morphological rule application. Machine Translation, 18(4) (2004) 275-297.

[14] Marquez, L., Padro, L. and Rodriguez, H., A machine learning approach to POS tagging. Machine Learning, 39(1) (2000) 59-91.

[15] Rani, P., Pudi, V. and Sharma, D.M., A semi-supervised associative classification method for POS tagging. International Journal of Data Science and Analytics, 1(2) (2016) 123-136.

[16] van Halteren, H. and Rem, M., Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters. *Language resources and evaluation*, *47*(4) (2013) 1233-1259.

[17] Petrochenkov, V.V. and Kazennikov, A.O., A statistical tagger for morphological tagging of Russian language texts. Automation and Remote Control, 74(10) (2013) 1724-1732.

[18] Dawa, I., Aishan, W. and Dorjiceren, B., Design and Analysis of a POS Tag Multilingual Dictionary for Mongolian. IERI Procedia, 7 (2014) 102-112.

[19] Das, B.R., Sahoo, S., Panda, C.S. and Patnaik, S., Part of speech tagging in Odia using support vector machine. Procedia Computer Science, 48(2015) 507-512.

[20] Ptaszynski, M. and Momouchi, Y., 2012. Part-of-speech tagger for Ainu language based on higher order Hidden Markov Model. Expert Systems With Applications, 39(14) (2012) 11576-11582.

[21] Pecheux, N., Wisniewski, G. and Yvon, F., Reassessing the value of resources for cross-lingual transfer of POS tagging models. Language Resources and Evaluation, 51(4) (2017) 927-960.

[22] Alhasan, A. and Al-Taani, A.T., POS tagging for arabic text using bee colony algorithm. Procedia computer science, 142 (2018) 158-165.

[23] Khan, W., Daud, A., Nasir, J.A., Amjad, T., Arafat, S., Aljohani, N. and Alotaibi, F.S., Urdu part of speech tagging using conditional random fields. Language Resources and Evaluation, 53(3) (2019) 331-362.

[24] Schulz, S. and Ketschik, N., From 0 to 10 million annotated words: part-of-speech tagging for Middle High German. Language Resources and Evaluation, 53(4) (2019) 837-863.

[25] Magistry, P., Ligozat, A.L. and Rosset, S., Exploiting languages proximity for part-of-speech tagging of three French regional languages. Language Resources and Evaluation, 53(4) (2019) 865-888.

[26] Myint, S.T.Y. and Sinha, G.R., Disambiguation using joint entropy in part of speech of written Myanmar text. International Journal of Information Technology, 11(4) (2019) 667-675.

[27] Mohammed, S., Using machine learning to build POS tagger for under-resourced language: the case of Somali. International Journal of Information Technology, 12 (2020) 717-729.

[28] de Oliveira, L.F.A., e Oliveira, L.E.S., Gumiel, Y.B., Carvalho, D.R. and Moro, C.M.C., Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. Research on Biomedical Engineering, 36(3) (2020) 267-276.

[29] Akhil, K.K., Rajimol, R. and Anoop, VS, Parts-of-speech tagging for malayalam using deep learning techniques. International Journal of Information Technology, 12(3) (2020) 741-748.

[30] Feng, X., Feng, Z., Zhao, W., Qin, B. and Liu, T., 2020. Enhanced Neural Machine Translation by Joint Decoding with Word and POS-tagging Sequences. Mobile Networks and Applications, 25(5) (20201722-1728.

[31] Besharati, S., Veisi, H., Darzi, A. and Saravani, S.H.H., 2021. A hybrid statistical and deep learning based technique for Persian part of speech tagging. Iran Journal of Computer Science, 4(1) (2021) 35-43.

[32] Maulana, A. and Romadhony, A., Domain Adaptation for Part-of-Speech Tagging of Indonesian Text Using Affix Information. Procedia Computer Science, 179 (2021) 640-647.