POS Tagging for Amharic Text: A Machine Learning Approach

SINTAYEHU HIRPSSA[1], G.S. LEHAL[2]

[1, 2] Department of Computer Science,
Punjabi University, India
sintsha2002@gmail.com; gslehal@gmail.com

Abstract: In this paper, our focus is the problem of automatic prediction of Parts of Speech tagging of words in Amharic language text. We conducted a comparison between statistical-based taggers. These are Conditional Random Field (CRF), an HMM-based Trigrams'n'Tags (TnT) Tagger, and Naive Bays (NB) based tagger. We compare the performances of all taggers with the same size of training and testing dataset. Also, various types of language-dependent and independent feature set have been formed, and on each tagger a combination of them are applied. As experimental result revealed that CRF based tagger has achieved best performance than others. The best accuracy obtained from our experiment using CRF is 94.08%. The study has also shown that linguistic features play a decisive role for minimizing or handling the challenges that stems from the morphological complexity of the language.

Keywords. Natural language processing, Feature set, Machine Learning, Conditional Random Fields, Amharic Language

(Received November 21, 2019, accepted June 9, 2020)

1. INTRODUCTION

A natural language's syntax and structure, like Amharic, are linked to a set of specific rules, conventions and principles that determine how words are combined into phrases, combine phrases into clauses, and combine clauses into sentences [1]. These rules and conventions are a key ingredient while developing natural language processing tools. To apply human language technology, we must use any of the following techniques as a pre-processor; Part-of-Speech (POS) tagging, Chunking or Parsing. In this paper, our focus is on Amharic POS tagging. POS tagging is used to find the lexical categories of each word in a sentence or to convey semantic information based on the syntactic context of the word [2], [3].

Analysing POS of single words in a text is a challenging task because terms can have various tag categories as they are used in a different context and some speeches may complex or unspoken [4]. POS tagging is used as a pre-processor component widely in information extraction (IE), voice synthesis, Name Entity Recognizer (NER), Word Sense Disambiguation (WSD), and Text to Voice (TTS) [5].

Amharic is a major language spoken in Ethiopia and is part of the Afro-Asian super family's Semitic branch [1]. There are two peculiar issues, which greatly affects the implementation of Amharic NLP applications. The first one is the lack of adequate resources and tools; however, recently, through the availability of the tagged corpus [6] and Amharic morphology analyser in public [7], Amharic has made some major steps forward. The second issue that perpetuates the problem rigorously is that the language is morphologically too rich, as the number of non-vocabulary words are usually large. Although such bottlenecks are rampant there, researchers are trying to put their unremitted effort for developing high-performance POS tagger for the Amharic language since 2001, though they are very few. Based on this, the aim of this study is also another exertion to improve the accuracy of Amharic POS taggers implemented so far. Thus, we conduct experiments via applying state-of-the-art tagging

methods, so to find out which method shows superior performance for this rich morphological language.

2. TRENDS IN AMHARIC POS TAGGING

As the best knowledge of the researcher, solely three Amharic POS taggers have attempted to develop using different machine learning methods range from the classifier to sequence labelling models. The efforts of POS tagging in Amharic started in 2001s [8]. The model was implemented using HMM and he compiled 25 POS tags for the first time, which extracted from a page long text, which was served as a groundwork tag for the following researchers. Also, the experiment was conducted by one-page long corpus as training and testing dataset and reported 87% accuracy. While taking, the size of the corpus he used, into consideration, it is difficult to conclude that the accuracy achieved was reasonable. Henceforth, different NLP researchers, those native and non-speakers have shown enthusiastic interest in the language to develop a POS tagging model [1], [9], [10].

Adafre [9] have explored the use of CRF for Amharic POS tagging. He collected five news articles 15,000 entries with their POS tags (Noun, Verb, Adjectives, and Adverb) and manually annotated them, which then used as training and testing set. The proposed tagger achieved 74% F-measure. A result of the experiments proves that Amharic taggers performance gets to an advanced degree once the comprehensive linguistic resources can be available. He extremely claimed that huge amount of annotated data is required to meet a performance which will comparable with the state-of-the-art results in other morphologically rich languages like [5], [11], Mohamed & Kubler [12]. Due to this fact, Getachew & Demeke [12] then took duty for developing Amharic tagged corpus. They collected 1065 news articles that contain 210K tokens from Walta Information Center (WIC), a private news agency in Addis Ababa, and then they tagged manually by 31 POS tags. Next, Gamback et al. [1] proposed Amharic POS tagging, they trained three Amharic POS tagger models and compared their performance each other. They carried out experiments using, frameworks such as TnT, SVM and MaxEntthey, and a corpus developed by [12]. Each tagger provided an effective accuracy; however, TnT was superior among all with an accuracy of 88%.

Afterwards, Binyam [12] came with a quite successful tagger, where he used several taggers. The researcher performed the experiments using Brill, CRF models, SVM, TnT models, and using manually crafted features. In this work, each proposed system had evaluated in a corpus which was used by Gamback et al. [1]. The CRF based model has outperformed than others by yielding an accuracy of 90.54%. He claimed that the size of the training data, the quality of the tag set, and feature set are the major ingredients which greatly affect the performance of the models. Therefore, we came to recognize that the POS tagging task not exclusively depends upon the dataset accustomed in the training phase of the model, but also, the feature set and tag set used is equally important. Generally, though few POS tagger was carried out for Amharic, continues exertion is required until a tagger come to exist that have human-level accuracy.

3. AMHARIC POS TAGS

Typical having a benchmark word class is the first matter before POS tagger development, due to this, now a day's two major public tag-sets in English, brown corpus has 87 tags and Penn Treebank has 48 tags [14]. Contrary, there is no accepted and default POS tag-set in Amharic, this makes don't trivial to develop Amharic POS tagger easily. The first tag set for POS tagger was compiled by Getachew [8], he designs 25 tags, Nonetheless, if we face paucity of the training dataset, i.e., lack of annotated corpus, it is better to use a few POS tags, since large POS tags may amplify the tendency of making the data-sparse. Thus Adafer [9] revised prior Amharic tag set, reduced into ten major tags during his experimentation. However, these tags capture the abstractive class of words, they have a great deal of limitation, which is a hierarchical relational detail would not be indicated clearly. A hierarchical tag-set, organized into major classes and subclasses seem to be a preferred design strategy. As a result of this, Demeke & Getachew [6] have derived, 31 classes of tags, from WIC news corpus, but basic classes are only eleven: N, PRON, ADJ, ADV, V, PREP, CONJ, INT, PUNC, NUM and UNC. And these basic classes are further divided based on the type of word only: don't comprise information on grammatical groups. Binyam [10] stated that although the tag-set has a tag for nouns with a preposition (NP), with the conjunction (NC) and with both prepositions and conjunction (NPC), it does not receive a separate tag for proper and plural nouns. Therefore, such nouns are assigned tag N and PRON a tag for pronouns that are not joined with others. We must agree with this notion and

used this tag-set (32 tags) for our experiment. Once the POS tag-set has been analyzed and compiled, the next step is to look at automatic methods of examining Amharic words, which we widen in the subsequent sections.

4. METHODOLOGY

POS tagging can be performed either using linguistically induced rules, which applied to distinguish the tag ambiguity [13]; or by a probabilistic method [14], [15], which uses statistical models and a corpus, choosing the tag order which maximizes the product of the token probability and tag sequence probability [14]. Besides, a combination of these approaches can be used for POS tagger development, known as a hybrid approach. The approach of this work is categorized under the statistical approach. More specifically, in our experiment, three machine learning-based taggers will be used, for the purpose of comparison and signifying the most effective for POS tagging in Amharic. These are CRFs from discriminative and Navies Bays (NB) based multinomial and Hidden Markov model-based Trigrams'n'Tags (TnT) from the generative model category. We enlighten below why we select each model for POS tagging comparison in Amharic.

CRFs is very common in sequential labelling problem and has shown achievement in POS tagging even in morphologically rich languages [5], [11], [12], [16]. This model has a great to handle a feature set to understand the characteristics of words, this led to tag words successfully. The idea of CRFs in POS tagging is creating *P(y|x)*model for the input sequence of words $X = x1, x2, \ldots, xn,$ and a label sequence $Y = y1, y2, ., yn,$ where yibelongs to the set of POS tags [14]. The model's label sequence$Y$ has the highest probability of all tag sequences in the output word sequence X. As such, CRF does not require to plainly model $P(x)$ and depending on the task, might so yield higher accuracy, in part since they require few parameters to be learned. Hence CRFs framework is often more suitable when complex and overlapping features are employed [15], [17]. The second tagger that we used is TnT, is a stochastic tagger, based on HMM [14], a proficient tagger that could be trained in various languages. It analyzes the sequential history of word-tag pairings in a specified sentence using the Markov Method [18].

TnT is based on tri-gram analysis (depends on two preceding tags) and language-independent statistical POS tagger [14]. Also, it includes a means of smoothing and handling unknown words. The Smoothing process is executed using linear interpolation, means the weight is determined by deleting interpolation, so we can easily train for the Amharic. TnT has a tendency to use supplementary features in training, for example, capitalization and suffixes. This then recognized where tags are mistakenly assigned, and tries to induce the correct rules through different context-sensitive templates. Then, it re-tags the dataset according to patterns learnt. Furthermore, the unique characteristic of this tagger is, the tags of unknown words are predicted based on the word suffix alone [14].

Another algorithm assessed in this study is Naive Bays based tagger, here the POS tagging process is considered as a classification problem. NB is also a successful algorithm in the classification task, also have the capability to make a parameter estimation based on a limited amount of training data. As well, it showed promising results in POS tagging for morphologically rich languages too [23], [24].

We used the NLTK and Sklearn Python library, which provides a set of computational linguistics and NLP program modules. Both allow us to develop numerous NLP applications by providing various algorithms for implementation of the POS tagger such as HMM-based TnT, memory-based tagger (MBT), n-gram; and classifier algorithm such as multinomial, SVM, DT Multinomial, etc.

5. Experiment and Results

5.1. Corpus preparation

The main components utmost necessary to develop an accurate and consistent POS tagger using statistical approach is quantitative and qualitative training data [2]. Because machine learning algorithms often require to be trained on huge volumes of tagged data to yield a successful model. However, Amharic is under resource language as stated by different NLP scholars, this is a major hampering that makes Amharic POS tagger unable to accomplish advanced performance like as other rich languages, Nevertheless, we tried to get large enough corpus compiled from two sources: the first is from ELRC which have ~ 210K token, and manually tagged with 31 tags by Demeke & Getachew [6].

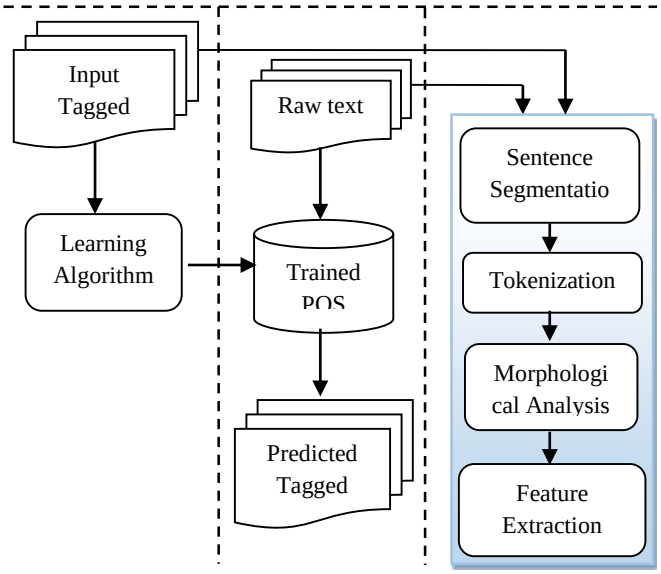

Figure 1: Conceptual architecture of an Amharic POS framework

The second corpus is a religious corpus [19] containing 116 K tokens which manually tagged with 62 tags. Regarding the tag set as we mention both corpora have a different number of tags so that we have converted into 31 benchmark tags. Various inconsistencies which must be cleaned in the entire corpus has been observed. For instance, compound word got a single POS tag, in another place different tags are assigned to decomposed words. E.g., in Amharic "4 netib 6" (4 point 6) is assigned a tag "NUMCR", in another place a tag "NUMCR" is assigned for 4, 6 and assign "PUNK" for neTb (Point) separately. Furthermore, "hakim'bet" (Hospital), assigned <N> and separate POS tags are assigned in another place. To solve this problem, the expressions were tokenized on space and given the tag that together they had in the first place. Then we listed word and tag pair per line, to make appropriate for the learners. Last but not list, inconsistency problem is related to tokens receiving multiple tags under the same conditions. A thorough attempt has been constituted to identify and correct most of them. After all this activity, the corpus has come 16451 sentences (~321,109 tokens).

5.2. Feature sets

As Figure 1 illustrates feature extraction is the major component after pre-processing tasks have been conducted. Most of NLP algorithm demands rigorous features to develop an effective system. Essentially, the feature set enables the algorithms to train faster, reduces the complexity of the system, and makes it easier to interpret. Features also help to disambiguate the words to some extent as well as it increases the accuracy of tagging when a vague word is encountered. A crucial aspect of feature-based probabilistic modelling is to obtain the appropriate facts about the data [2]. In our experiments, different feature both language dependent and independent have been examined based on the different possible grouping of available words and tag context; and prove the most suitable set of features for Amharic POS tagging in each framework. Most of them were used by Binyam [10] so we adopted with a little modification.

- ***Contextual feature:*** as of other's language, Amharic is also being suffered from the problem of word ambiguity. To find a resolution for such ambiguities, we defined a feature set in a context window of the current word, $\{w_{i-2}, w_{i-1}, - w_i, w_{i+1}, w_{i+2}\}$. As various study reviled the surrounding words can play a vital role in deciding the POS tag of the current word.
- ***Morphological feature:*** In the POS tagging process many words, those unseen during learning, would be tagged arbitrarily by the model because there are no context-based events within the model to facilitate correct tagging. Discovering prefix/suffix of each word gives a way to decrease irrational tagging of a word. As Molina [20] stipulates that, the prefixes and suffixes of the current word as features are found effective for highly inflected languages like Amharic. As we tend to delineate in Figure 1, once tokenization has taken place, the morphology analyser instrument [7] take the position to analyse the morphology of the input word discretely.
- ***Lexical features:*** A lexicon in Amharic has used to improve the performance of the POS tagger further. The lexicon incorporates the root words and basic information about the POS. Unknown words are searched in the lexicon if there is a match, then POS tags obtained from the lexicon are assigned to the words. The most frequently happening POS label is also allocated with a word in the training corpus.
- ***Orthographic features:*** it encodes information about the type of text that appears in the token. Some of the orthographic features:

- *Made up of digits:* a binary-valued feature and used to check whether the current token consists of only digits or not. It helps to find the numerical expressions, particularly used for the quantifier number tags. Also, we used a feature which analyses whether the word made from both alphabet and numeric.
- *Contains symbol:* a binary-valued feature has been incorporated to check whether the current token contains any special symbol (e.g., %, $ etc.). This feature helps to recognize symbols tags.

All aforementioned features are given to the algorithm in the form of a template and then, for example, CRF creates a model assigning feature weights to the individual features. After the model is created, a test data set is used to the POS tags of the test example. The entire process can be modelled as a combination of different functions. The ultimate model for predicting POS tags is specified in Equation 1. (as discussed in see Jurafsky and Martin [21]

Model = x (t (f (p(x; y)))) ----------(1)

p, position between word x and y

f, feature Representation for the words x, y

t, transfer of features from x to y

x, model edifice using the suggested features

The new word's POS tag is then calculated by the model based on the token's adjacent tags.

5.3. Evaluation and Discussion

Multiple experiments have been conducted to determine the appropriate features for POS tagging in Amharic. And evaluated using 10-fold cross-validation to get a significant accuracy and dependable result. At the initial experiment, we implement the baseline tagger. Here we assumed the tagger simply predicts the maximum probable class based on the class probabilities learned from training data, also based on the frequent POS tag observed from the training data.

All taggers have been trained with the same training set, similarly, the same test set is used to evaluate all taggers, this allows us to compare the results directly and helps to fairly identify which tagger is outperformed others. As the experimental result revealed, the CRF achieved 94.08% F-measure, where this performs is better than the performance achieved by TnT- 87.39 % F-measure and NB- 81.25 % F-measure. However, the TnT is the second-best tagger here, it has shown some improvements when compared with the tagger that developed using a similar model by Gamback [1] and Binyam [10]. Also, CRF-based tagger has shown better performed on tagging unknown words, achieved 76.44% F-measure, whereas TnT-based tagger achieved 61.28% and NB-based classifier achieved 41.85%. In all cases, the NB-based tagger shows the least performance. Practically, this tagger's limitation arisen from the prediction means once the tagger is learned and then if it's going to incorrectly predict for the POS for the current word, the next word could have a high probability to be predicted incorrectly.

Regarding the features, we evaluate each feature independently and in a combination thoroughly, starting from a baseline experiment. The baseline experiment in CRF performed well, however, isn't remarkable as compared with other latest CRF based POS taggers in litrature. Therefore, we continued our experiment by applying a different feature set until the best accuracy has been obtained. Continuing the experiment by combinig a baseline tagger with a morphological feature of each word, to analyze the influence of affix. The morphological feature has enhanced the baseline overall F-measure by 5.29 points.

From this, we can understand that the POS-tagging performance can be dropped by 5.29 points, because of the nature of language that is rich in morphology. Moreover, baseline with contextual feature-based tagger demonstrated 18.26 pints exceeded an overall F-measure while compared to the baseline model. Additionally, this feature combination revealed the best performance that exceeds by 12.97 points than a combination of baseline with morphology, implying that contextual information is a key for POS tagging in the Amharic using CRF framework.

Model	Feature set used	F1 (%)
CRF	Baseline model	68.35
	Baseline + Context(word and POS tag)	86.61
	Baseline + Morphological	73.64
	Baseline + Contextual + Morphological	90.58
	Baseline + Contextual + Morphological + + Position	91.14
	Baseline+ Contextual + Morphological + Position + Orthographic	94.08
NB	Baseline model	62.12
	Baseline + Context(word and POS tag)	75.54
	Baseline + Morphological	67.27
	Baseline + Contextual + Morphological	79.33
	Baseline + Contextual + Morphological + + Position	79.14
	Baseline+ Contextual + Morphological + Position + Orthographic	81.25
TnT	Contextual	87.39
	Contextual + Suffix	86.17

Table 1: Overall performance of the taggers in terms of average F-measure

Next, only 0.56 pints improvement has been observed while combining a features such as baseline, morphological and contextual features over the addition of the positional feature. Thereby demonstrating that the position feature does not affect the performance of tagger significantly. In the end, it is perceptible that 2.94 point increments on the performance of the tagger while we combined the orthographic feature with baseline, contextual, morphological and positional features. This indicates that adding the orthographic feature of each word is necessarily important for word's class identification. Particularly, some of the orthographic features such as 'is_symbole' do not contribute to improving the performance, but "is_digit" and "is_punk" feature shows an invaluable improvement in overall performance. In brief, at the combination of features with certain parameters, a maximum accuracy has been found by CRF-based tagger. These features are the context words [-2, 2]; up to 4 characters for prefix 3 characters for suffix of the current word; two previous word and one POS tag; and orthography (binary-valued including "is_digit" and "is_punctuation") feature.

In general, the experimental result showed a contextual feature takes a major role in all algorithm for POS tagging model development in Amharic. Next to context feature, analyzing the morphology of words play a crucial role in improving the tagging performance. Moreover, CRF demands the linguistic features higly, this also lets to word disambiguation when the ambiguous word is encountered and so then increase the accuracy of Amharic POS tagging. Lastly, the experiment revealed that "building effective features is extremely necessary for building a sequential learning model" and also gives us a notion that the tagger performance would be boosted up if feature set tuning could be done rigorously.

Going further, it has also been investigating the effect of the data size and learning capability of the algorithms. We made experiments with gradually increasing the training size by 40k each turn, starting from 80k till to 320K. Our experiment attested that increasing the size of training data led to better results in all models, though, the degree of performance improvement varies. Among all models, TnT is highly affected by the amount of training set, implying that its performance greatly depends on the training data size. TnT has gained a total of 14.89 on the overall F--measure, however, 75% of this gain is recorded between 80K to 120K, which is the peak improvement. This finding is in line with Tachebelie [22]. Contrariwise, CRF-based tagger's performance has not affected significantly by the size of the training data, but highly influenced by features. CRF-based tagger gained a total of 4.48 performance improvement on overall accuracy. The size of training data affects the classification performance of NB based classifier, it shows a linear increments on performance.Generally, it can be said that there is a clear correlation between the training data size and the performance of ML algorithms; the larger the training data size, the better the performance.

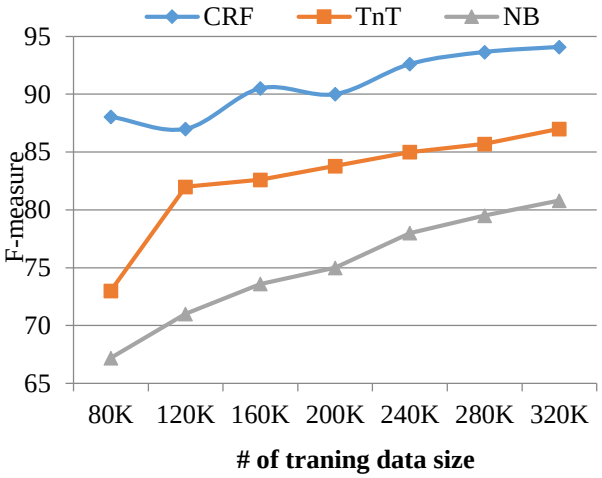

Figure 2: The data size versus performance with the best feature set

Generally, our experiment gives us a notion that the accuracy of the POS tagging task must boost up if feature set identification could be done perfectly. Finally, our CRF based achieved astonishing performance which appears as much better than any other tagger that developed for the Amharic language so far.

6. CONCLUSION

The POS taggers described here are very straightforward and competent in automatic tagging for Amharic text. Besides, we have discussed the exact nature of various features such as lexical, morphological, contextual and orthographic. And several experiments have been carried out to comprehend the best features set on each tagger, and also a technique for handling unknown words. The result of our experiment showed

that the CRF model is a super tagging strategy for Amharic languages, as the accuracy of the tagger is less affected, after it reaches at some point, as the amount of training data increases compared with other methods. Although it is highly affected by the amount and type of features set and can be improved its accuracy. We believe that the future enhancements of this work would be, shifting the approach into deep learning, which is convenient for automatic feature exploitation and representation from a raw text.

REFERENCES

[1] B. Gambäck, F. Olsson, A. A. Argaw, and L. Asker, "Methods for Amharic part-of-speech tagging," in *First Workshop on Language Technologies for African Languages*, 2009, p. 104–111.

[2] H. Tseng, D. Jurafsky, and C. Manning, "Morphological features help POS tagging of unknown words across language varieties," in *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005, p. 32–39.

[3] J. H. M. Daniel Jurafsky, "Part-of-Speech Tagging," *Speech Lang. Process.*, vol. 1, p. 1–28, 2019.

[4] M. Banko and R. C. Moore, "Part of speech tagging in context," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, p. 556–56.

[5] F. Albogamy and A. Ramsay, "Fast and robust POS tagger for Arabic tweets using agreement-based bootstrapping," in *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016, p. 1500–1506.

[6] G. Demeke and M. Getachew, "Manual annotation of Amharic news items with part-of-speech tags and its challenges," *Ethiop. Lang. Res. Cent. Work. Pap.*, vol. 2, p. 1–16, 2006.

[7] M. Gasser, "HornMorpho: a System for Morphological Processing of Amharic, Oromo, and Tigrinya," in *Conference on Human Language Technology for Development*, 2011, p. 1–55.

[8] Mesfin Getachew, "part-of-speech tagging," Addis Ababa University, 2001.

[9] S. F. Adafre, "Part of speech tagging for Amharic using conditional random fields," in *Proceedings ACL Workshop on computational approaches to Semitic langEstimation of Conditional Probabilities With Deciuages*, 2005, p. 47–54.

[10] G. G. Binyam, B. Gebrekidan, and G. G. Binyam, "Part of Speech Tagging for Amharic," in *Proceedings of Natural language processing & human language technology*, 2010, p. 1–20.

[11] S. P. K. Gadde and M. V. Yeleti, "Improving statistical POS tagging using Linguistic feature for Hindi and Telugu Improving statistical POS tagging using linguistic features for Hindi and Telugu," in *International Conference on Natural Language Processing*, 2008.

[12] S. Kübler and E. Mohamed, "Part of speech tagging for Arabic," *Nat. Lang. Eng.*, vol. 18, no. 4, p. 521–548, 2012,

[13] E. Brill, "Rule-based part of speech," in *Conference on Applied Natural Language Processing*, 1992, p. 152–155.

[14] T. Brants, "TnT: a statistical part-of-speech tagger," in *Proceedings of the sixth conference on Applied natural language processing*, 2000, p. 224–231.

[15] H. Schmid and F. Laws, "Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging," in *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008, p. 777–784.

[16] N. X. Bach, N. D. Linh, and T. M. Phuong, "An empirical study on POS tagging for Vietnamese social media text," *Comput. Speech Lang.*, vol. 50, p. 1–15, 2018.

[17] W. E. I. Jiang, X. Wang, and Y. I. Guan, "Improving sequence tagging using machine-learning," in *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, 2006, p. 13–16.

[18] Z. Ghahramani, "An Introduction to Hidden Markov Models and Bayeian Networks," *Appl. Comput. Vis.*, p. 9–42, 2001.

[19] G. Ibrahim and S. HL, "Machine Learning Approaches for Amharic Parts-of-speech Tagging," in *Proc. of ICON*, 2018, p. 74–79.

[20] L. C. Molina, L. Belanche, A. Nebot, À. Nebot, J. Girona, and C. Nord, "Feature Selection Algorithms: A Survey and Experimental Evaluation," in *International Conference on Data Mining.*, 2002, p. 306–313.

[21] P. Jane Austen, "Sequence Processing with Recurrent Neural Networks," in *Speech and Language Processing*, D. J. & J. H. Martin, Ed. 2019.

[22] Y. T. Martha, T. A. Solomon, and L. Besacier, "Part-of-Speech Tagging for Under-Resourced and Morphologically Rich Languages – The Case of Amharic," in *Conference on Human Language Technology for Development*, 2011, p. 50–55.

[23] Amey K. Shet T and Surabhi N. "Text Classification using Naïve Bayes, VSM and Pos Tagger," *International Journal of Ethics in Engineering & Management Education*, Vol. 4, no.1, p.66-73, 2017 .

[24] Rund M, Nazlia O and Omaia A, "Arabic part of speech tagging using k-nearest Neighbour and

Naive Bayes classifiers combination," *Journal of computer science*, Vol. 10, no. 9, p.1865-1873, 2017.