Informing Science:
the International Journal of
an Emerging Transdiscipline

An Official Publication
of the Informing Science Institute
InformingScience.org

Inform.nu

Volume 26, 2023

ANALYSIS OF MACHINE-BASED LEARNING ALGORITHM USED IN NAMED ENTITY RECOGNITION

Francis Mithanga Kamau*	Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya	mithash41@gmail.com
Kennedy O. Ogada	Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya	kenogada@gmail.com
Cheruiyot W. Kipruto	Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya	wilchery68@gmail.com

* Corresponding author

ABSTRACT

Aim/Purpose	The amount of information published has increased dramatically due to the information explosion. The issue of managing information as it expands at this rate lies in the development of information extraction technology that can turn unstructured data into organized data that is understandable and controllable by computers
Background	The primary goal of named entity recognition (NER) is to extract named entities from amorphous materials and place them in pre-defined semantic classes.
Methodology	In our work, we analyze various machine learning algorithms and implement K-NN which has been widely used in machine learning and remains one of the most popular methods to classify data.
Contribution	To the researchers' best knowledge, no published study has presented Named entity recognition for the Kikuyu language using a machine learning algorithm. This research will fill this gap by recognizing entities in the Kikuyu language.
Findings	An evaluation was done by testing precision, recall, and F-measure. The experiment results demonstrate that using K-NN is effective in classification performance.
Recommendations for Researchers	With enough training data, researchers could perform an experiment and check the learning curve with accuracy that compares to state of art NER.

Accepting Editor Eli Cohen | Received: October 31, 2022 | Revised: January 8, January 15, 2023 | Accepted: January 16, 2023.
Cite as: Kamau, F. M., Ogada, K. O., & Kipruto, C. W. (2023). Analysis of machine-based learning algorithm used in named entity recognition. *Informing Science: The International Journal of an Emerging Transdiscipline, 26,* 69-84. https://doi.org/10.28945/5073

Impact on Society	NER helps recognize important textual components including names of individuals, places, and monetary value among others. When dealing with enormous datasets, it is critical to sort unstructured data and to find vital information by identifying the major entities in a text.
Future Research	Future studies may be done using unsupervised and semi-supervised learning algorithms for other resource-scarce languages.
Keywords	named entity recognition, memory-based learning algorithms, semantic web problems, K-Nearest Neighbor (KNN), precision, recall, F-Score measures

INTRODUCTION

The word named in the expression named entity as defined by Kripke (1972) restricts the task to entities for which one or many rigid designators stand for the referent. The term 'named entity' was first mentioned in the Sixth Message Understanding Conference (MUC-6) as the task of identifying the names of all the people, organizations, and geographic locations in a text, as well as time, currency, and percentage expressions. Since the MUC-6, other events have been devoted to named entity recognition (NER), for example, IREX (Demartini et al., 2010) and TREC Entity Track (Balog et al., 2010).

The definition of named entity (NE) follows a classification of NEs into two main categories: generic entities, e.g., person, organization, and geographical location, and domain-specific entities, e.g., genes and terms. The development of NER models has been necessitated by the increased amount of data to process (Mikhailov & Shavrina, 2020). The spread of fake news in social media and various other media which is a threat to social and national peace has also led to research on detecting fake news as true or false by use of NLP for textual analysis (Khanam et al., 2021). This paper's main objective is to analyze memory-based learning algorithms and use them to develop a framework to identify named entities.

LITERATURE REVIEW

WHAT IS NAMED ENTITY RECOGNITION?

Named entity recognition categorizes entities such as people's names, locations, organization, time, currency, percentage expressions, and currency from a given text. Named entity recognition is used in information extraction where texts or entities are identified in a non-structured way. The selection of a tag set in the NER task has been a challenge that leads to the extraction of limited types of entities, such as people, organization, temporary expressions, and numeric expressions. The introduction of the GENIA corpus, dedicated to entity types such as RNA, cell, DNA, and protein, made studies easier. Studies have also recognized drugs, disease symptom names, and chemical names (Goyal et al., 2018).

CHALLENGES IN NER

Several challenges are encountered when creating a robust NER. These challenges are ambiguity in text and the availability of resources and nested entities. An ambiguous text appears in one place as a named entity and another as a common noun or is used to refer to a different entity. For example, Jordan refers to a person's name and place. The unavailability of a large corpus and gazetteers is challenging when implementing NER systems. Some languages like Hindi, Urdu, and Punjab are resource-poor, making implementing the NER task challenging. Nested entities are inside other named entities, which is hard to recognize and therefore requires segmenting and labeling (D. N. Shah & Bhadka, 2017).

Applications of NER

NER is an important foundation in managing enormous amounts of digital information stored in an unstructured and structured form. Computing has made advances, and humans will be able to communicate in spoken language with computers. Natural language processing (NLP) enables computers to understand and derive meaning from natural language text by identifying entities. Information extraction has applied NER, where proper names and named entities carry essential information about the text itself; therefore, the accuracy of information extraction depends on them (Gangadharan & Gupta, 2020). Extraction of named entities improves semantic search making it more robust. Identifying and evaluating a query search makes search engines meet users' intent (Kostakos, 2020). NER build a system that answer a question asked in natural language by human beings. Factoid-type questions usually start with a wh-word (Who, Which, Where, What) and require answers in a small sentence or a phrase. Incorporating NER in a QA system makes finding answers to some questions easy.

Conversion of text or speech from the source language to the language target using a computer without human involvement requires proper name identification. Correctly identifying them impacts the translation's overall syntactic structure and local context. Automatic Named Recognition Systems improve machine translation quality (Temnikova et al., 2019). Building a knowledge base or ontologies requires extracting concepts and entities from data and learning from semantics; therefore, it needs support from NER. An example of this tool is KnowItAll (Rodríguez-García et al., 2021). People make opinions freely on various topics on social media, and many make decisions based on their views. OPINE is an example of such a system for extracting attributes and opinions (Gupta & Agrawal, 2020).

Techniques Used in Named Entity Recognition

Techniques for NER are largely classified into three categories: rule-based approaches, learning-based approaches, and hybrid approaches (Goyal et al., 2018).

Rule-based approaches

Earlier NER systems were mostly based on handcrafted rules. These systems used information lists such as gazetteers, a book with a described list of places, and rule-based syntactic-lexical patterns to detect and categorize entities (Flores & Pinto, 2020). By employing specific language domain features, they obtain significant accuracy. Some limitations of the rule-based method are that they are expensive to create and maintain, they are domain-specific, and they are not portable. In addition to this, they require human expertise in language knowledge and programming skills. Hence rule-based cannot be used across different language domains, which shifts researchers' interest to machine learning approaches. Rule-based approaches are seen in papers such as Alfred et al. (2013) and Drovo et al. (2019). Rule-based is applied in the Bengali language where the researcher noted the difficulty of maintaining the NER since its language dependent. In rule-based learning, few entities follow specific patterns, such as dates, emails, and time. The rule-based approach has better accuracy for entities with patterns than the machine-learning method. The challenge for the rule-based approach is that it suffers from language dependence limitations.

Machine learning approaches

A machine learning-based named entity recognition framework aims to transform identification problems into classification problems. The problem is then solved using a statistical model (Anandika & Mishra, 2019). NER uses machine learning algorithms to compute relationships and patterns in the text. Machine learning algorithms are advantageous over rule-based since they are trainable, can adapt to other domains, and are less expensive to maintain trained data. State-of-the-art NER generally implements machine learning algorithms based on statistical machine learning. Machine learning (ML) techniques are broadly classified into three categories: supervised learning, semi-supervised learning, and unsupervised learning (Daud et al., 2017).

Hidden Markov Model

HMM is a generative type of sequence-based model and works in three phases. First is the annotation phase, which produces tagged documents from raw corpus or text. The second is the training phase, where parameters are set. These parameters in HMM are three: first is the phase called Start Probability, denoted as π; the second is Emission Probability (B), and the third is Transmission Probability (A). The last phase is testing, where the user gives the test sentence to the framework. Based on the previously computed parameters of the HMM using the Viterbi algorithm, the optimal state sequence for the test sentence is given (Lay & Cho, 2019). Based on the research conducted by Syachrul et al. (2019) on Indonesian language Qur'an translations using the Hidden Markov Model resulted in the highest F1 of 76% found after combining features that are pre-processing and POS tag.

However, the lowest result of 46% is found using only pre-processing. For future works, the author proposes more development to the research to prevent name ambiguity. For example, the name Allah is often changed to Him in the Quran. A study on Indonesian medicinal plants resulted in the lowest F1 of 41% and the highest F1 of 72%. The study introduces names, substances, places, and uses of Indonesian medical plants. Research by Lay and Cho (2019) for Myanmar, a language spoken by the natives majority of Burmans, developed a named entity using HMM. The training data of 3000 sentences and testing data of 150 sentences of the Myanmar language were used. The evaluation was done by checking the accuracy, Precision, Recall, and F-Measure. The results of the experiment are in Table 1.

Table 1. Myanmar NER results using HMM

Measure	Results
Accuracy	0.9523
Precision	0.9928
Recall	0.9523
F-Measure *(harmonic mean of the precision and recall)*	0.9721

The shortcomings of Myanmar NER using HMM were a lack of proper resources due to a lack of ready corpus and a lack of capitalization in the language, unlike some languages, for example, English.

Conditional random field (CRF)

CRF is a sequence modeling algorithm that assumes features are dependent on each other and considers future observations when learning new patterns. It combines the advantages of MEMM and HMM. It incorporates dependent features and context-dependent learning. CRFs define a conditional probability $p(y|x)$ over label sequences given a particular observation sequence x. These models allow the labeling of an arbitrary sequence X by choosing the label sequence y' that maximizes the conditional probability $p(y'|x')$ (Freire et al., 2012). Research that has used CRF includes Freire et al. (2012) in the identification of three entities which are person, location, and organization, form poorly structured data in bibliographic contents using the CRF approach and attained a maximum precision of 91% at 55% recall and a maximum recall of 82% at 77% precision. Chen et al. (2015) used an annotated text on clinical data. They created a NER that identified three entities: treatment, problem, and test using a supervised CRF approach, and got an F-score of 80%. Majumder et al. (2012) researched a named entity using the CRF approach to identify drug names and disease names from a diagnosis discussion forum. An annotated dataset of 100,000 words was trained, and 12,000 words were considered for testing. The features set used included affixes, capitalization, and word features. Large unannotated data raised the recall by 2% and the F-score value by 1.1%, respectively.

In addition, Munarko et al. (2018) developed a NER for Indonesian tweets using the CRF classifier. The researcher used 8,000 tweets in the model by grouping them into formal and informal tweets. The test was done with training with 2,000 training data. The result of the model was a Recall and Precision of 62% and 87% for formal tweets. For a mixed corpus of formal and informal tweets, the result was a precision of 60% and a recall of 86%, respectively. The result was measured in ten-fold cross-validation.

Related recent research on CRF is a Named entity recognition using CRF (Patil et al., 2020). The researchers developed NER for the Marathi language using 27,177 sentences. The sentences were manually tagged and used to train and evaluate the system. Satisfactory performance was achieved. The challenge of the model was the inflected (words modified to have different grammatical categories) nature of the Marathi language and rich morphology. Techniques, such as Stemming and Lemmatization, can be studied to handle inflection in Marathi text.

Maximum Entropy Model

The principle of Maximum Entropy is that the best probability model for the data is the one that maximizes entropy over the set of probability distributions that are consistent with the evidence. In his task, Jung (2012) performed an extraction of NER on micro-tasks that stream online on social media sites. The researcher used the Maximum Entropy (ME) approach and resolved the challenge of small-size text by merging associated micro texts. These associations include semantic, social, and temporal associations. Semantic looks at the similarity of word features, and social association looks at the digital ID of the user. In contrast, temporal association checks the closeness between two micro texts about time. The results showed a high accuracy of 90.3% using a micro text cluster.

Support Vector Machine (SVM)

SVM is a supervised type of machine learning algorithm. Unlike neural networks that only look for dividing hyperplane in an instance, SVM finds the best hyperplane in fed space. Some of the NER using the SVM approach are named entities (Ekbal et al., 2012), which identified four entities: person, location, organization, and miscellaneous using Hindi and Bengali. Hindi experiment results were a precision of 90.22%, recall of 89.41%, and F-score of 89.81%. Bengali had a precision of 91.65%, recall of 91.66%, and F-score of 91.65%. The authors conclude that for resource-scarce languages such as Hindi and Bengali, SVM Produced satisfactory results. The other example of a model that implemented SVM is by Yusup et al. (2019). The authors in their papers identified the following features: title case, which identifies letters starting with capital letters.

Naïve Bayes (NB)

Some of the work done using Naive Bayes includes research by Azalia et al. (2019), where the authors used the Naïve Bayes Classifier to create a Name Index translation of Hadith in the Indonesian language. The IOB Tag was done for the dataset with a precision of 28.52%, recall of 33.5%, and F-score of 31.5 %. Using morphological features, Titlecase produced 47.84%. For POS Tag and Unigram, both Lexical features, POS Tag produced a better result of 76.75% compared to unigram, which had a 71.41% F1 score. The study combined Unigram, POS Tag, and Title Case to achieve an F1 of 82.63%. These results showed that the more features are used, the better the performance.

Granik and Mesyura (2017) carried out research to detect fake news using naive Bayes classifier. The authors created a software system using this method and tested it on a group of news posts from Facebook. The system was able to correctly identify about 74% of fake news on the test set. The paper also highlighted some ways to improve the accuracy. Their findings indicated that artificial intelligence can be used to tackle the problem of detecting fake news.

Decision Tree learning

According to Elçi et al. (2020), decision tree learning is widely used because it can explicitly and visually express rules. It can be represented as a set of If-Then Rules. Decision trees are constructed in a top-down manner where the feature with the most information about the target label assumes the root of the decision tree. This classifier is in the form of a tree structure, and every node is represented as a leaf. This principle is followed when forming the tree. The entropy measures the purity or the impurity of data. When data is non-homogenous to the target label, it is said to be pure. It is impure when it is a mixture of target variables. Entropy is zero for impure data and high for pure data (Sarker, 2021). Information Gain and Gini Index are two procedures of attribute selection. There are three commonly used algorithms in decision tree learning: Classification and Regression Tree (CART), Iterative Dichotomiser3 (ID3), and C4.5. The advantage of ID3 is that it reduces repeated implementation of operations (Anandika & Mishra, 2019).

In Al-Hegami et al.'s (2017) research on biomedical named entity recognition, the researchers used 10-fold cross-validation to gain an unbiased evaluation of the system performance. Compared with KNN, the results were lower in Precision, Recall, and F-measure. It means feature sets have a smaller impact on decision tree classification performance than on KNN classification.

K-Nearest Neighbor Algorithm

K-Nearest Neighbor (K-NN) is a Supervised Machine Learning algorithm. It is one of the simplest algorithms used for regression and classification problems. It is non-parametric which means it does not assume underlying data. Simple yet very useful with distinguished performance, the K-NNN technique has been widely utilized in data mining and machine learning. After training sample data, classification is used to forecast the labels of test data points. Although many different categorization techniques have been presented by scholars over the past few decades, K-NN remains one of the most often used techniques (Pandey & Jain, 2017). Some of the research done using the K-NN algorithm are sentiment analysis on Twitter using SVM and K-NN where the results indicated that K-NN performed better than SVM (Rezwanul et al., 2017). With 40% less training data than the vanilla NER model, KNN-NER can produce results that are equivalent (Wang et al., 2022).

Hybrid model

A hybrid model is another approach to the development of the NER. In this approach, the hybrid approach finds results by combining handcrafted rules with machine learning methods or by combining two or more machine learning approaches. In a review by Goyal et al. (2018), the authors proposed a hybrid named entity recognition for the Chinese language. The authors detected three entities, namely organization, person, and location using a CRF model whereby the entities are labeled with tags including O, E, I, B2, and B1. The authors noticed the results were not satisfactory and therefore implemented post-processing in the subsequent step, which included some transformation-based learning and rules. The author archives better results with the O, E, I, and B tag set of labels. The result takes less time and fewer resources from the system for training. A result F-score of 93.49% is achieved.

An examples hybrid model is Yang et al. (2017) for opinion mining and sentiment analysis using a Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT), where SVM performs well for sentences with simple structure but poor performance for complicated sentences. The GBDT performs well for long sentences with many words. The researcher combined the two algorithms using the stacking approach, which is an ensemble learning approach that combines two learning algorithms. The other researchers (Santoso et al., 2020) use Hybrid Conditional Random Fields and K-Means to recognize entities in Indonesian news documents. The experiments performed with the proposed hybrid model produced the best result of 87.18%.

Research by Drovo et al. (2019) that merged the rule-based and Hidden Markov models indicates that the machine learning method was used to classify entities and the rule-based approach to increase accuracy. The researchers noted that the Bengali language, unlike English, is resource-scarce, and there was not much research done on the language and named entity recognition. The corpus was tagged from an available online Bengali newspaper called Prothom Alo. The corpus had 10,000k words annotated with seven tags. The tags were person, location, organization, mail, date, and other. The evaluation was done for the 690 sentences that included 10,000 words and was done in two ways. The test results were an f-score of 68.98% for the first way, which was done rule-based and Hidden Markov Model simultaneously. Table 2 shows a summary comparison of strengths and weakness of various machine learning-based algorithms.

Table 2. Comparison of different machine learning-based algorithms

Machine learning-based method algorithm	Advantage	Disadvantage
Support vector machine	High accuracy for small and good training dataset	Needs time to train the dataset
K-NN	Needs no training data Simple to implement Robust to the noisy training data More effective if the training data is large	Determining value of k may be complex
CRF and HMM	Reduced human effort on maintaining rules	Preparing annotated data
Hybrid	Higher performance	Dependent on a combination of different feature selection methods. Complicated architecture
Decision Tree	No need to normalize data. Pre-processing step requires less time to code	Its mathematical calculation requires more Memory

Source: (Juárez-Orozco et al., 2018).

After a literature review analyzing various machine learning algorithms, this paper implements the K-Nearest Neighbor (KNN) Algorithm.

EXPERIMENTS AND RESULTS

The taggings are as follows. I-PER represents the name of the person in the NER. NNE represents none required entity in the NER. B-ORG stands for the name of an organization, while B is the beginning of that name. I-ORG stands for any word that follows from the beginning of the noun phrase. I-PLAC represents the name of a location. B-TIME represents the beginning of the time noun phrase, while I- TIME represents any word that follows from the beginning of the noun phrase of the NER. I-MON represents the name of the month. The <utt> is a sentence delimiter that identifies the end of a statement to the memory-based tagger. Table 3 is an annotated corpus for kikuyu language while Table 4 is a dictionary for data validation. Figure 1 is sample algorithm input and output graphical user interface.

Table 3. Sample annotated corpus for Kikuyu language

Name	Description
Ibuku	NNE
ria	NNE
uhoro	NNE
wigii	NNE
utuuro	NNE
wa	NNE
Jesu	I-PER
Kristo	I-PER
,	NNE
muru	NNE
wa	NNE
Daudi	I-PER
,	NNE
muru	NNE
wa	NNE
Iburahimu	I-PER
:	NNE
Iburahimu	I-PER
agituika	NNE
ithe	NNE
wa	NNE
Isaaka	I-PER

Source: (Sang & De Meulder, 2003).

Table 4. Sample data dictionary for validation

Name	Class tag	Description	Field size	Data type
Ibuku	NNE	Non-name Entity	255	String
ria	NNE	Non-name Entity	255	String
Jesu	I-PER	Person	255	String
Kristo	I-PER	Person	255	String
,	NNE	Non-name Entity	255	String
Karatina	B-ORG	Organization	255	String
Nairobi	B-ORG	Organization	255	String
Geneal	I-ORG	Organization	255	String
Hospital	I-ORG	Organization	255	String
Ikumi	NUM	Number	255	String
Ithano	NUM	Number	255	String

Source: (Goyal, 2021).

Below is a sample of Kikuyu sentences with the output expected from the NER.

- Kimani arikia githomo giake athiire cukuru wa Ngoima High School
- Tukwanjia igeranio cia muico wa mwaka mweri wa Kanyuahungu cukuru wa Kimathi University
- Kimathi oimia mburi githaa kia Ime riatika
- Kibirigwi Primary School ikoragwo mwena wa Gitunduti

Output

- Kimani(I-PER) arikia(NNE) githomo(NNE) giake(NNE) athiire(NNE) cukuru(NNE) wa(NNE) Ngoima(B-ORG) High(I-ORG) School(I-ORG)
- Tukwanjia(NNE) igeranio(NNE) cia(NNE) muico(NNE) wa(NNE) mwaka(NNE) mweri(NNE) wa(NNE) Kanyuahungu(I-MON) cukuru(NNE) wa(NNE) Kimathi(B-ORG) University(I-ORG)
- Kimathi(I-PER) oimia(NNE) mburi(NNE) githaa(NNE) kia(NNE) Ime(B-TIME) ri-atika(ITIME)
- Kibirigwi(B-ORG) Primary(I-ORG) School(I-ORG) ikoragwo(NNE) mwena(NNE) wa(NNE) Gitunduti(I-PLAC)

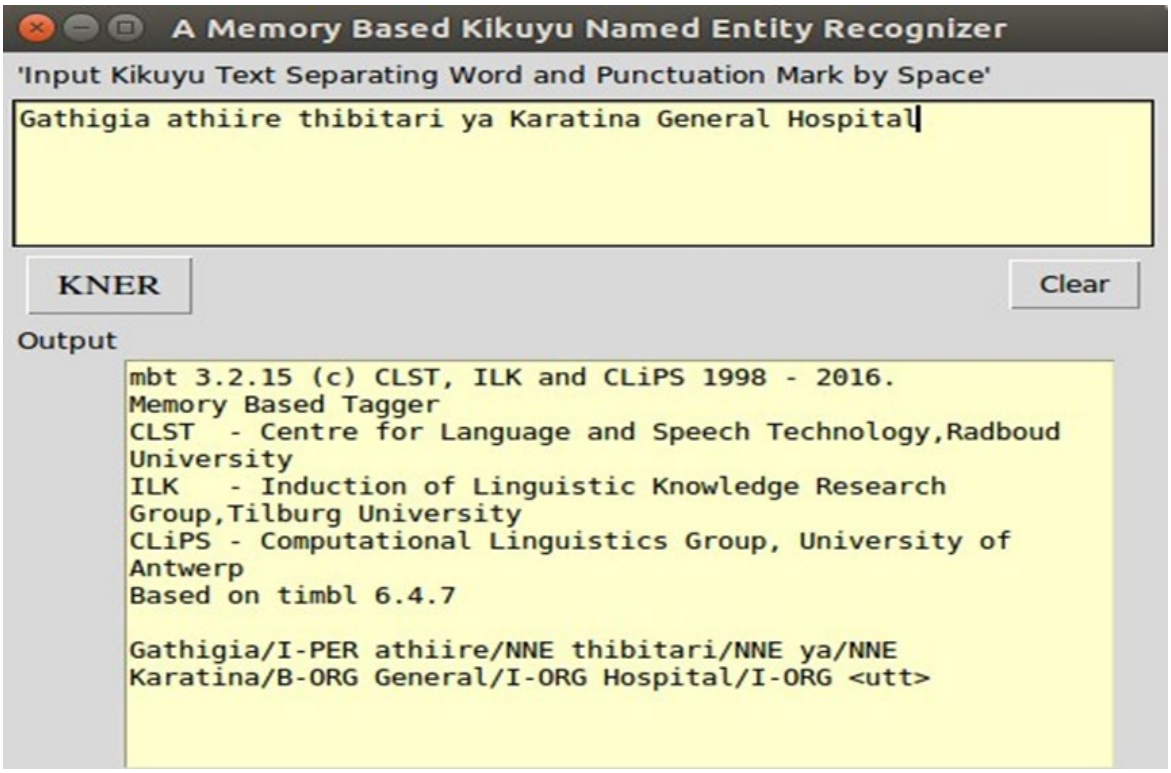

Figure 1: Sample algorithm for the Kikuyu language designed through this study
(Goyal, 2021)

EVALUATION METRICS

In this section, we describe the experimental results obtained based on a corpus of 17,000 words. The evaluation measured Precision, Recall, and F-score, which are calculated based on true positives (TP), false positives (FP), and false negatives (FN). True positives are the correctly labeled instances. False positives are the incorrectly labeled instances, and false negatives are the missed-out instances by the framework. F-score is the weighted mean of Precision and Recall. These metrics are formulated as follows:

Equation 1: Recall

$$Recall = \frac{Number\ of\ instances\ correctly\ labelled}{Total\ number\ of\ relevant\ instance\ labeled}$$

Equation 2: Precision

$$Precision = \frac{Number\ of\ instances\ correctly\ labelled}{Total\ number\ of\ instance\ labeled}$$

Equation 3: F Score

$$F - score\ = 2 \times \frac{Precision\ .Recall}{Precison + Recall}$$

Table 5 is a Sample dictionary for evaluations.

Table 5. Sample dictionary for evaluations

Class tags	Description	Data Type	Field size
NNE	Non-named Entity	String	255
I-PER	Name of a person	String	255
I-PLACE	Name of a Place	String	255
I-TIME	Time	String	255
B-ORG	Organization	String	255
I-ORG	Organization	String	255
I-MON	Month	String	255
B-TIME	Time	String	255

Source: (Loomis, 2021).

Experiment Results

We used a balanced evaluation, known as the cross-validation technique, based on the principle that all but one of the dataset's n chunks should be used to model the framework. This procedure is done "k" times, hence the name "k-fold cross-validation." A different portion is kept for testing in each iteration. By averaging the outcomes of each cycle, the ultimate score is determined. For NER tasks, 10-fold cross-validation is frequently utilized (Chen et al., 2015; Liu & Zhou, 2013). Data is summarized, and details are extracted from the 1st to the 10th fold. The different folds are used for cross-validation. This allowed error correction when a fold had language mistakes. Table 6 shows different class tags and the percentage precision, recall and F-score measure.

Table 6. Evaluation of different classes in various folds

Class tags	Precision	Recall	F-Score
NNE (Non-Named Entity)	0.9939	0.9943	0.9942
I-PER (Person Name)	0.9618	0.9692	0.9655
I-PLACE (Place)	0.8571	0.8108	0.8333
I-TIME (Time)	1.0000	0.6000	0.7500
B-ORG (Organization)	0.9000	0.6428	0.7723
I-ORG (Organization)	0.9647	0.9534	0.9590

Source: (D. N. Shah & Bhadka, 2017).

For this NER, the class B-ORG with 0.64 and I-TIME with 0.60 had a low recall. These are rare classes, and the lack of enough data is a bottleneck that contributed to a low F-core. Other classes had good recall and precision, which resulted in better performance of the F-Measure. Table 7 indicates various folds percentage accuracy.

Table 7. Number of words used versus overall accuracy

Number of words	Accuracy %
1700	85.79
3400	89.24
5100	91.23
6800	92.98
8500	94.55
10200	95.25
11900	96.34
13600	96.59
15300	97.81
17000	98.49

Source: (D. N. Shah & Bhadka, 2017).

Figure 2 indicates that an increased number of words has an increased accuracy percentage, which means accuracy is directly proportional to the data available in the corpus.

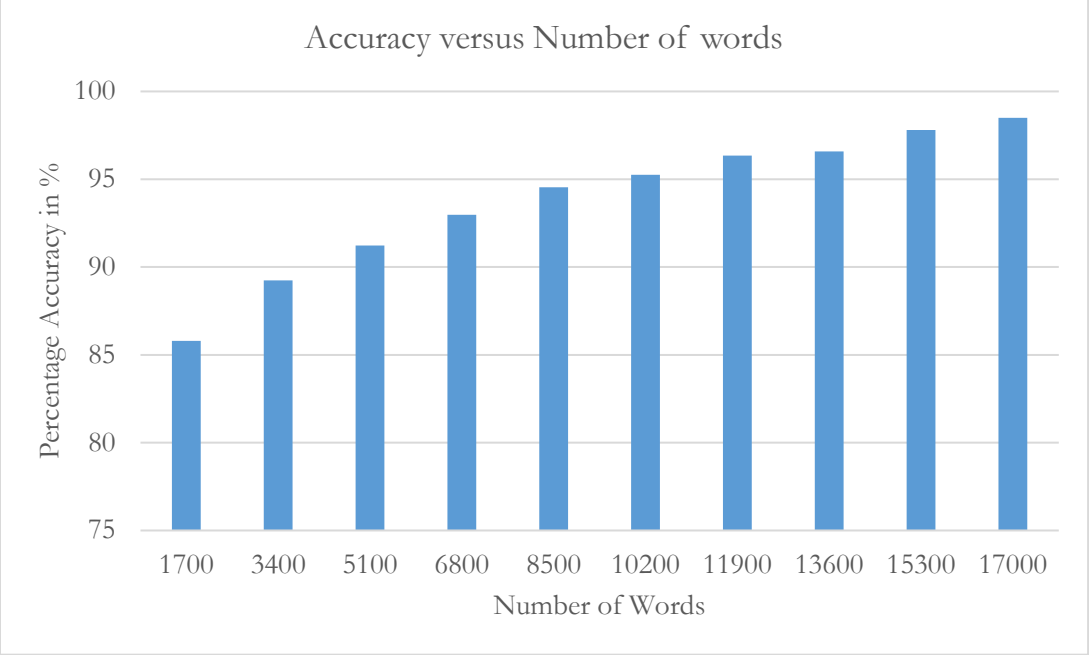

Figure 2. Accuracy versus number of words used

Table 8 shows how different folds have distinct precision percentages depending on the words put in the classifier. For instance, the morphological arrangement of the letters is not the same for every named entity. Recall gave the number of times a class was correctly predicted divided by the number of times a class appears in the dataset. The Recall of the NER was equally good, with an average of 84.50%. Except for handling B-type tags as in fold 9, which had 50%, and fold five, with 72.32%. The low percentage is due to scarce data to handle this rare class effectively.

Table 8. Precision, recall, and F-score for the 10 folds

Folds	Precision (in %)	Recall (in %)	F Score (in %)
Fold1	93.764	92.489	93.104
Fold2	96.151	95.696	95.622
Fold3	99.911	99.242	99.307
Fold4	97.623	82.275	94.322
Fold5	93.013	72.324	95.003
Fold6	95.222	93.754	94.468
Fold7	94.649	93.267	93.775
Fold8	94.796	77.694	90.183
Fold9	79.279	50.932	92.714
Fold10	92.108	87.352	83.076
Average	93.64	84.50	93.16

Source: (D. N. Shah & Bhadka, 2017).

When the framework is compared to R. Shah et al.'s (2010) work on Named entity recognition for the Swahili language, with the best F1 score of 81.5%, and Kikamba NER, which on average by analyzing unknown and known entities got 96.42% having used a larger corpus of 2,7754 words. The research gave a satisfactory average of 93.16%, which was acceptable considering Kikuyu is a resource-scarce language.

CONCLUSION

There is high demand for NER due to the exponential growth of the internet, with a wealth of information available to people. In addition, there is a deep penetration of connected smart mobile devices. It is difficult to find an automated solution for a linguistic problem. Identifying structural and grammatical rule relationships and their dependencies is a difficult endeavor. In this paper, we analyzed the machine learning algorithm and implemented K-NN-based NER to recognize Kikuyu-named entities. We achieved a satisfactory result.

FURTHER RESEARCH

Large, annotated corpora are the most important prerequisite for supervised machine learning training and testing methods. Still, they are difficult to come by in many resource-constrained languages, particularly in Africa. On the other hand, semi-supervised and unsupervised algorithms require less or no annotated data. Future studies may be done using unsupervised and semi-supervised learning algorithms for other resource-scarce languages. With enough training data, researchers could experiment and check the learning curve with accuracy that compares to state of art NER.

Funding and conflicts of interests/competing interests

This research is not funded, and the author (Francis Mithanga Kamau) has the approval of his supervisors to publish in this journal.

REFERENCES

Alfred, R., Leong, L. C., On, C. K., Anthony, P., Fun, T. S., Razali, M. N. B., & Hijazi, M. H. (2013). A rule-based named-entity recognition for Malay articles. In H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, & W. Wang (Eds.), *Advanced data mining and application* (pp. 288–299). Springer. https://doi.org/10.1007/978-3-642-53914-5_25

Al-Hegami, A. S., Farea Othman, A. M., & Bagash, F. T. (2017). A biomedical named entity recognition using machine learning classifiers and rich feature set. *International Journal of Computer Science and Network Security*, *17*(1), 170–176. http://pper.ijcsns.org/07_book/201701/20170126.pdf

Anandika, A., & Mishra, S. P. (2019, May). A study on machine learning approaches for named entity recognition. *Proceedings of the International Conference on Applied Machine Learning, Bhubaneswar, India,* 153–159. https://doi.org/10.1109/ICAML48257.2019.00037

Azalia, F. Y., Bijaksana, M. A., & Huda, A. F. (2019). Name indexing in Indonesian translation of Hadith using named entity recognition with naïve Bayes classifier. *Procedia Computer Science*, *157*, 142–149. https://doi.org/10.1016/j.procs.2019.08.151

Balog, K., Serdyukov, P., & de Vries, A. (2010, November). Overview of the TREC 2010 Entity Track. *Proceedings of the 19th Text Retrieval Conference, Gaithersburg, Maryland.* https://www.researchgate.net/publication/221037684

Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., & Xu, H. (2015). A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, *58*, 11–18. https://doi.org/10.1016/j.jbi.2015.09.010

Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: A survey. *Artificial Intelligence Review*, *47*(3), 279–311. https://doi.org/10.1007/s10462-016-9482-x

Demartini, G., Iofciu, T., & de Vries, A. P. (2010). Overview of the INEX 2009 Entity Ranking Track. In S. Geva, J. Kamps, & A. Trotman (Eds.), *Focused Retrieval and Evaluation* (pp. 254–264). Springer. https://doi.org/10.1007/978-3-642-14556-8_26

Drovo, M. D., Chowdhury, M., Uday, S. I., & Das, A. K. (2019, June). Named entity recognition in Bengali text using merged hidden Markov model and rule base approach. *Proceedings of the 7th International Conference on Smart Computing & Communications*, *Sarawak, Malaysia,* 1–5. https://doi.org/10.1109/ICSCC.2019.8843661

Ekbal, A., Saha, S., & Singh, D. (2012, August). Active machine learning technique for named entity recognition. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Chennai, India,* 180–186. https://doi.org/10.1145/2345396.2345427

Elçi, A., Sa, P. K., Modi, C. N., Olague, G., Sahoo, M. N., & Bakshi, S. (Eds.). (2020). *Smart computing paradigms: New progresses and challenges.* Springer. https://doi.org/10.1007/978-981-13-9680-9

Flores, O. R., & Pinto, D. (2020). Proposal for named entities recognition and classification (NERC) and the automatic generation of rules on Mexican news. *Computacion y Sistemas*, *24*(2), 533–538. https://doi.org/10.13053/CyS-24-2-3377

Freire, N., Borbinha, J., & Calado, P. (2012). An approach for named entity recognition in poorly structured data. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, & V. Presutti, V. (Eds.) *The semantic web: Research and applications* (pp. 718–732). Springer. https://doi.org/10.1007/978-3-642-30284-8_55

Gangadharan, V., & Gupta, D. (2020). Recognizing named entities in agriculture documents using LDA based topic modelling techniques, *Procedia Computer Science*, *171*, 1337–1345. https://doi.org/10.1016/j.procs.2020.04.143

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review, 29*, 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001

Goyal, C. (2021, June 23). Named Entity Recognition | Guide to Master NLP (Part 10). *Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON),* 900–903. https://doi.org/10.1109/UKRCON.2017.8100379.

Gupta, N., & Agrawal, R. (2020). Application and techniques of opinion mining. In S. Bhattacharyya, V. Snášel, D. Gupta, & A. Khanna (Eds.), *Hybrid computational intelligence*: *Challenges and applications* (pp. 1–23). Academic Press. https://doi.org/10.1016/b978-0-12-818699-2.00001-9

Juárez-Orozco, L., Martinez Manzanera, O., Nesterov, S., Kajander, S., & Knuuti, J. (2018). The machine learning horizon in cardiac hybrid imaging. *European Journal of Hybrid Imaging*, *2*, 1–15. https://doi.org/10.1186/s41824-018-0033-3

Jung, J. J. (2012). Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, *39*(9), 8066–8070. https://doi.org/10.1016/j.eswa.2012.01.136.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012040. https://doi.org/10.1088/1757-899x/1099/1/012040

Kostakos, P. (2020, December). Strings and things: A semantic search engine for news quotes using named entity recognition. *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, The Hague, Netherlands*, 835–839. https://doi.org/10.1109/ASONAM49781.2020.9381383

Kripke, S. A. (1972). Naming and necessity. In D. Davidson, & G. Harman (Eds.), *Semantics of natural language* (pp. 253–355). Springer. https://doi.org/10.1007/978-94-010-2557-7_9

Lay, K. K., & Cho, A. (2019). Myanmar named entity recognition with Hidden Markov Model. *International Journal of Trend in Scientific Research and Development*, *3*(4), 1144–1147. https://doi.org/10.31142/ijtsrd24012

Liu, X., & Zhou, M. (2013). Two-stage NER for tweets with clustering. *Information Processing and Management*, *49*(1), 264–273. https://doi.org/10.1016/j.ipm.2012.05.006

Loomis, C. (2021, January 6). nerman: Named Entity Recognition System Built on AllenNLP and Optuna. *Optuna*. https://medium.com/optuna/nerman-named-entity-recognition-system-built-on-allennlp-and-optuna-c044c319b955

Majumder, M., Barman, U., Prasad, R., Saurabh, K., & Saha, S. K. (2012). A novel technique for name identification from homeopathy diagnosis discussion forum. *Procedia Technology*, *6*, 379–386. https://doi.org/10.1016/j.protcy.2012.10.045

Mikhailov, V., & Shavrina, T. (2020). *Domain-transferable method for named entity recognition task*. arXivLabs. https://doi.org/10.48550/arXiv.2011.12170

Munarko, Y., Sutrisno, M. S., Mahardika, W. A. I., Nuryasin, I., & Azhar, Y. (2018). Named entity recognition model for Indonesian tweet using CRF classifier. *IOP Conference Series: Materials Science and Engineering*, *403*, 012067. https://doi.org/10.1088/1757-899X/403/1/012067

Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, *9*(11), 36–42. https://doi.org/10.5815/ijcnis.2017.11.04

Patil, N., Patil, A., & Pawar, B. V. (2020). Named entity recognition using conditional random fields. *Procedia Computer Science*, *167*, 1181–1188. https://doi.org/10.1016/j.procs.2020.03.431

Rezwanul, M., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, *8*(6), 19–25. https://doi.org/10.14569/ijacsa.2017.080603

Rodríguez-García, M. Á., García-Sánchez, F., & Valencia-García, R. (2021). Knowledge-based system for crop pests and diseases recognition. *Electronics*, *10*(8), 905. https://doi.org/10.3390/electronics10080905

Sang, E. F. T. K., & De Meulder, F. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition* (arXiv:cs/0306050). arXiv. http://arxiv.org/abs/cs/0306050

Santoso, J., Setiawan, E. I., Yuniarno, E. M., Hariadi, M., & Purnomo, M. (2020). Hybrid conditional random fields and K-Means for named entity recognition on Indonesian news documents. *International Journal of Intelligent Engineering & Systems*, *13*(3), 233-245. https://doi.org/10.22266/ijies2020.0630.22

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science*, *2*, 1–21. https://doi.org/10.1007/s42979-021-00592-x

Shah, D. N., & Bhadka, H. (2017). A survey on various approach used in named entity recognition for Indian languages. *International Journal of Computer Applications, 167*(1), 11–18. https://doi.org/10.5120/ijca2017913878

Shah, R., Lin, B., Gershman, A., & Frederking, R. E. (2010). SYNERGY: A named entity recognition system for resource-scarce languages such as Swahili using online machine translation. *Proceedings of the Second Workshop on African Language Technology,* 21-26. https://www.semanticscholar.org/paper/SYNERGY-%3A-A-Named-Entity-Recognition-System-for-as-Shah-Lin/bce8977b41f359454f21af801ae7e0338b41760a

Syachrul, R. M. M.A.K., Bijaksana, M. A., & Huda, A. F. (2019). Person entity recognition for the Indonesian Qur'an translation with the approach Hidden Markov Model-Viterbi. *Procedia Computer Science*, *157*, 214–220. https://doi.org/10.1016/j.procs.2019.08.160

Temnikova, I., Orasan, C., Corpas Pastor, G., & Mitkov, R. (2019, September). Preface. *Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology*, *Varna, Bulgaria* https://aclanthology.org/W19-8700.pdf

Wang, S., Li, X., Meng, Y., Zhang, T., Ouyang, R., Li, J., & Wang, G. (2022). *k-NN-NER: Named entity recognition with nearest neighbor search*. arXivLabs. http://arxiv.org/abs/2203.17103

Yang, K., Cai, Y., Huang, D., Li, J., Zhou, Z., & Lei, X. (2017, February). An effective hybrid model for opinion mining and sentiment analysis. *Proceedings of the IEEE International Conference on Big Data and Smart Computing, Jeju, South Korea,* 465–466. https://doi.org/10.1109/BIGCOMP.2017.7881759

Yusup, F. A., Bijaksana, M. A., & Huda, A. F. (2019). Narrator's name recognition with support vector machine for indexing Indonesian hadith translations. *Procedia Computer Science, 157*, 191–198. https://doi.org/10.1016/j.procs.2019.08.157

AUTHORS

Francis Mithanga Kamau is a student in Master of Computer Systems in Jomo Kenyatta University of Agriculture and Technology. He received his Diploma in Information Studies in KTTC, Nairobi, Kenya in 2010 then B.S. degrees in Library and Information Science from Kenyatta University of Kenya, 2014. Since 2015, he has been a Senior Library Assistant at Kirinyaga University, Library Department.

Dr. Kennedy Ogada is a Senior Lecturer at the Jomo Kenyatta University of Agriculture and Technology, School of Computing and Information Technology. His research interest is in machine learning and artificial intelligence.

Professor Cheruiyot W. Kipruto is a professor working for the Jomo Kenyatta University of Agriculture and Technology, School of Computing and Information Technology. His research areas include machine learning, multimedia systems and communications, information retrieval, image processing, semantic web, distributed databases and internet, data warehousing and theory of computation.

.