Global Scientific JOURNALS

GSJ: Volume 9, Issue 9, September 2021, Online: ISSN 2320-9186
www.globalscientificjournal.com

Using Classification Model for Information Extraction

Saleih Gero[1], Getu Girma [2]Shimels Hailu [3]

[1 &3] School of Computer Science Dire Dawa Institute Dire Dawa University

[2]School of Mechanical and Industrial Engineering Dire Dawa Institute of Technology

Dire Dawa, Ethiopia

Abstract

As the size of unstructured document containing relevant information become increase more and more, manually seeking for specific relevant information in this document is a difficult, tedious and time-consuming task. An information extraction system is a method that solves such difficulty by automatically extracting specific relevant information from such unstructured text documents and putting this information as structured pattern. In this study, an automated Afaan Oromoo Information Extraction System (AOIES) has been developed using supervised machine learning approach to extract the most relevant football news information from a collection of unstructured Afaan Oromoo sport news document. The Afaan Oromoo sport news documents used in this study was collected from Radio Fana Share Company Afaan Oromoo Department. To implement the AOIES, the tokenization, normalization, stop word removal language specific methods, the machine learning Naïve Bayes classification algorithm and various programming language tools are applied. The standard precision, recall and F-score evaluation metrics are used to evaluate the text classification and IE model accuracy of the developed system prototype. While experimenting the proposed model with training and testing dataset, the 10-fold cross validation method is applied. The developed system classification module achieved 91.7% and the IE model 94.6% F-scores performance by correctly predicting the instances. The above result indicates the developed system prototype has scored good performance by correctly predicting the instances using the Naïve Bayes algorithm. Generally, the evaluation result demonstrates that the machine learning classification algorithm can be adopted as information extraction method for the Afaan Oromoo text document.

***Key Words:*, Machine Learning, Naive Bayes, Afaan Oromoo , Information Extraction**

INTRODUCTION

Information extraction (IE) is an automated system that search and extract most interesting information including entities name, entities relation, and events from given documents. In another way, IE is the process of automatically detecting and then extracting the significant data from unstructured, semi structured or structured documents by using natural language processing and other modeling techniques. The automatic IE system process takes documents as an input and generate the respective structured information as a result and this generated information can be utilized by other users in different applications for further analysis purposes. Entities detection and extraction tasks are one of the commonest IE system tasks that identify basic entities name, date, time and some regular expressions. However, manually searching and extracting the particular relevant information which are existing under large size unstructured document is tedious and time-consuming task for users [1] [2]. To overcome this difficulty, a lot of researches have been conducted under the area of automatic IE system regarding to various domain specific tasks. Even though various automated IE systems have been developed for different languages such as Amharic, English and others countless to perform numerous domain specific tasks, it is difficult to apply these systems directly for Afaan Oromoo Language without some adoptions. Due to this reason, there is a need to have automated IE system for Afaan Oromoo language that extract certain valuable facts from a collection of unstructured Afaan Oromoo documents. Therefore, the objective of this study is to develop an automated IE system for Afaan Oromoo language that narrowing this gap. The developed Afaan Oromoo information system extracts the most relevant football news information from the Afaan Oromoo sport news document. The system is developed using supervised machine learning methods which consider the IE task as a classification problem point of view. A collection Afaan Oromoo sport news document collected from the Radio Fana Broadcasting Corporate is applied as training and testing dataset to develop the IE system. Generally, the contribution of this study is it demonstrate how to apply the automatic machine learning approach and NLP techniques to develop information system for Afaan Oromoo language and it predicts a good starting point for other researchers who need to participate in the related field with different domain specific researchable problem area [3].

LITERATURE REVIEW

Our country, Ethiopia is compassed of multination ethnic groups approximately with more than 80 nations. The state incorporates various language categories including Cushitic, Semitic, Omotic and Nilotic. The Afaan Oromoo language is one of the Cushitic language categories that is belonging to the Afro-Asiatic language family. The Afaan Oromoo language is a mother tongue for Oromoo people and it is spoken by the Oromoo people existing in the

Ethiopia, The Oromoo people are the largest ethnic group in in Ethiopia and they inhabited more than half percentage of population density in the country. Currently the Afaan Oromoo language has more than 60 million speakers in the country and it is commonly spoken in some neighboring regional states in the Ethiopia including Amhara, Benishangul Gumuz, Gambrella, SNNP, Afar and also in others neighboring east African countries such as Kenya, Uganda, Sudan, Djibouti, Eritrea and Somalia as well. As a result, the language is occupying the third level in Africa which is next to the Arabic and Hausa languages by the speakers of population density. The Afaan Oromoo language writing system is known as the "**Qubee**" which is adopted from the Latin- alphabet and currently this writing system is serving as language official scripting tool and communication means in the Oromia regional state government since 1991 E.C [29] [14] [31]**.**

Generally, a lot of research papers are publishing in the form of both hardcopy and softcopy versions with respect to various research scopes in the language domain and currently some media and educational institutions, social, political and religion organizations are using the Afaan Oromoo language as a means of scripting and communication tools.

As it was explained under the research work [14], the language contains 33 consonants in which 7 of them are the combination of these consonants. These 7 consonant letters are called **"QUBEE DACHAA"** and they include **CH, DH, NY, PH, SH, TS** and **ZH**. In the Afaan Oromoo alphabet the vowels are known by the name "**Dubbiftoota**". In the Afaan Oromoo language, these vowels can be appeared at the initial, central and end position of words. In the Afaan Oromoo alphabet there are five vowels known as short sound vowels which are written as '**a', 'e', 'o', 'u'** and '**i'**. Doubling these vowels may create another five basic vowels called long sound vowels which are written as **'aa', 'ee', 'ii', 'oo', 'uu'**. Hence, the Afaan Oromoo alphabet is consists of ten vowels categorized as 5 short sound vowels and 5 long sound vowels [31].

The machine learning and extraction module are the basic component while building IE systems using the selected appropriate strategy to handle the process of candidate entities recognition.

While using this approach, the knowledge engineers are not required and instead of this a set of training data along with small manpower and linguistic resources are required. This approach focuses on creating training data and only experts who trained with some domain specific of IE are needed and the information extraction system based on this approach are basically rely on automatically acquiring the extraction patterns. As explained by Getasew Tsedalu [15], the information extraction systems developed using this approach involve name identification and classification, full parsing or partial parsing, semantic classification of nominal mentions, coreference resolution, relation extraction and event extraction tasks. The benefit of automatic machine learning approach is it does not need more resources like skilled man power, money, time, it does not highly linguistic dependent to develop

since it learns automatically the patterns. Based on the above explanation and analysis, automatic machine learning approach is more convenient than knowledge engineering approach to develop Afaan Oromoo information extraction system.

MODEL DESIGNING METHODOLOGY

In order to train and model the proposed automated Afaan Oromoo Information Extraction System (AOIES), the machine learning method and NLP techniques are used. Just like others text information extraction systems, the basic architectural components of the proposed automated Afaan Oromoo Information Extraction System (AOIES) consist of the document browsing phase, data preprocessing phase, machine learning phase and postprocessing phase with respective unique subcomponents. The detail tasks of each subcomponents and utilized methods will discussed under the subsequent sections of this chapter.

Document Source and Corpus Development

For the sake of developing AOIES, the Afaan Oromoo sport news instances are collected from the Radio Fana Broadcasting Afaan Oromoo Service. The utilized techniques while collecting the news document include analyzing collected document and reviewing various related literatures. These collected sport news document reorganized and then used as a source of input text data to develop the proposed AOIES. Since the information extraction system is a language domain-specific task, these collected data needs to be annotated manually using similar procedures depending on name of candidate text type to be extracted. Accordingly, 500 football news items, 500 athletics news items and 500 other news items are utilized as a dataset to train and test the text classification module of the AOIES model. Similarly, 1800 football news instances are selected and used as a dataset to train and test the IE module of the proposed AOIES.

The AOIES Architectural Design

For the process of developing the AOIES, the supervised machine learning approach is applied and hence the IE system architectural design is constructed using the supervised machine learning algorithm. The AOIES architectural design is composed of four main phases called as the document browsing phase, preprocessing phase, machine learning phase and postprocessing phase as main components. The document browsing phase contains news data insertion module to handle the task of browsing Afaan Oromoo sport news document that provided as data to build the AOIES model, the preprocessing phase contains text Feature filtration module to handle the process of data cleaning task, the machine learning phase composed of the text classification and IE modules to control the process of learning predefined patterns and then extracting most relevant information with Naïve Bayes Multinomial algorithm, and the postprocessing phase is composed of information formatting module to format, store and filling the formatted information in to predefined output template form for display.

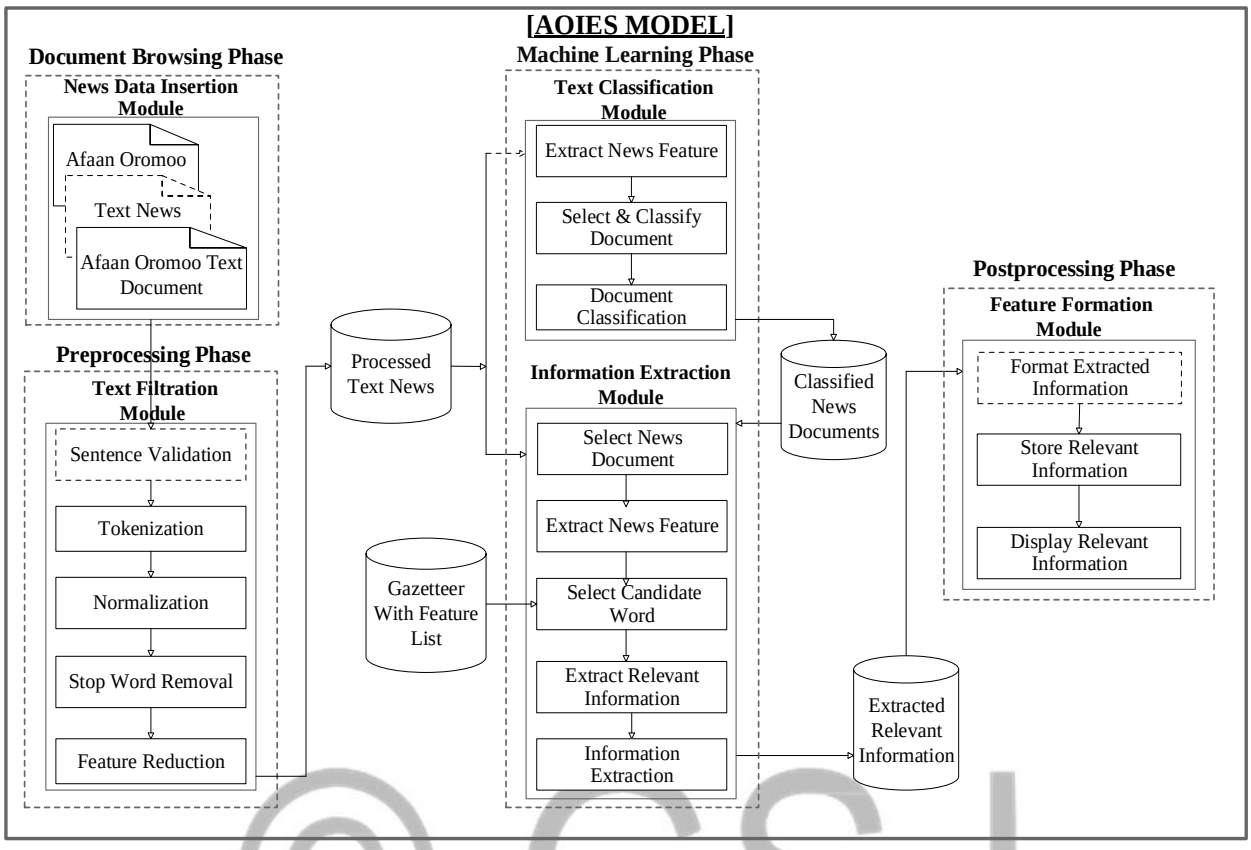

Figure 1- AOIES Architectural Design

Document Browsing Phase

The document is used as a basic input data for most natural languages processing tasks and the same fact is true for Afaan Oromoo Information extraction system. Like some other western languages such as English, Spanish and Portuguese which are based on the Latin alphabet writing system, the Afaan Oromoo language writing system also uses the Latin alphabet scripting with slight modification. In this study, the document browsing phase main task is to handle the task of browsing (loading) and accessing the Afaan Oromoo sport news document.

Preprocessing Phase

The preprocessing phase is a fundamental backbone for any NLP related task like IE. The preprocessing method used in this study is based on the feature filtering technique with the help of additional language specific rules to split the Afaan Oromoo sport news text document in to set of word segments. The final output of this phase is the preprocessed stream of words which will be utilized by the machine learning phase. This phase contains another subcomponent as feature filtration module which is consist of different subtasks categorized as sentences

validation, tokenization, normalization and stop word removal.

Machine Learning Phase

The AOIES machine learning phase is to learning a set of rules and syntaxes under the Afaan Oromoo language the training time. These rules and syntaxes are used to guide the machine learning module while classifying and extracting the relevant information from unstructured Afaan Oromoo sport news document. The machine learning module is the phase in which the real tasks of AOIES is developed from the Afaan Oromoo sport news training data. This phase is composed of the text classification and information extraction module as main modules which are explained bellow.

Text Classification Module

The main task of text classification module is to control the task of learning and classifying the Afaan Oromoo sport news document as football news, athletics news and others news with trained Naïve Bayes Multinomial classification algorithm. The existence and likeness of key term features are enables to identify the class categories of sport news. The text classification module output is serves as input data aimed at information extraction module. This module consisting of different submodules to handle task of extracting news feature, selecting relevant feature set that are used as attribute values to train and test the classification module. The various tasks of the text classification module include news feature extraction, news feature selection, news feature classification and text classification.

Information Extraction Module

After the Afaan Oromoo sport news document have been classified as football news, athletics news and unknown category news during the text classification module, the information extraction module selects only the football news class label, extract candidate feature set and select all the relevant candidate information with respect to the predefined candidate token categories. The process of detecting and extracting feature space representing the specified candidate tokens categories with respective relevant information is handled by specializing the machine learning algorithms used in the text classification module. The various tasks of the information extraction module include football news class selection, news information extraction, candidate tokens selection, relevant candidate information extraction with respect to all selected candidate tokens categories and the information extraction.

Post processing Phase

In the previous modules, the unstructured input document is processed in the preprocessing module, represented and the most relevant candidate tokens attributes are selected in the machine learning module based on their statistical notches. The postprocessing phase is the last phase of the AOIES and it contains feature formation module with various subtasks including format extracted information, storing extracted information and displaying the most relevant information accordingly. Hence, the main role of this phase is formatting the extracted football news candidate information, storing that information to the file rendering to the predefined output template and then making this relevant information ready to display accordingly. The final outputs of this phase are the structured Afaan Oromoo football news information categorized as game category, news publication year, news editor, matching teams, stadium, date-time of game, score of game and winner of game.

Result and Discussion

The information extraction evaluation results analysis is based on the accuracy percentage while predicting the football news candidate tokens with respect to all the predefined feature spaces using the Naïve Bayes extraction algorithm.

Detailed Accuracy Per Class

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.968	0.008	0.934	0.968	0.950	Game Category
0.926	0.005	0.951	0.926	0.938	News Publication Year
0.909	0.003	0.978	0.909	0.942	News Editor
0.979	0.008	0.939	0.979	0.958	Match Teams
0.929	0.003	0.968	0.929	0.948	Stadium of Game
0.946	0.008	0.928	0.946	0.937	Time of Game
0.956	0.012	0.920	0.956	0.938	Score of Game

0.957	0.005	0.961	0.957	0.959	Winner of Game
0.946	0.007	0.947	0.946	0.946	Average

Table 1-IE Detailed Accuracy Per Class

The overall accurate average result is referring to the generated final overall F-measure scores while detecting and extracting the football news candidate tokens correctly using the extraction algorithm. For example, looking to the last row of F-measure column in the above IE table 17 detailed accuracy per class, the accurate overall average F-measure result generated by Naïve Byes extraction algorithm is 94.6%. While comparing the proposed AOIES performance with other related IE works done using the supervised machine learning approach for other languages, the experimentation analysis result predicts the extraction model has been achieved fair performance with relative evaluation score. The overall final outcome of this study demonstrates it is possible to use the machine learning algorithm as information extraction task for the Afaan Oromoo language text document.

Contribution

The developed information extraction system is used to extract the relevant football news information automatically from unstructured Afaan Oromoo sport news text documents. The main contribution of this study can be summarized as follow. The Language specific tasks handling algorithms have been developed which enable the language speakers to easily extracts football news relevant information from the Afaan Oromoo sport news document. The extracted relevant information can be used as an input source by different news publication mass media institutes for further analysis in different application areas. It showed good starting point for other researchers to give attention on the language domain and it can be used as a standing backbone of further works in the domain and supporting reference material for other researchers who are working in similar domain problems. It proves that it is possible to adopt and use machine learning model for other related Afaan Oromoo Language based NLP tasks including information retrieval, text summarization, text classification, question and answering, text to speech synthesis, machine translation, speech recognition, parts of speech tagging and automatic sentence construction.

Conclusion

In this study, the automated Afaan Oromoo information extraction system (AOIES) has been developed using the Afaan Oromoo sport news document and supervised machine learning approach by incorporating language specific patterns. The developed system predicts and extracts the most relevant football news information from unstructured Afaan Oromoo sport news document using the Naïve Bayes classification learning algorithm with the integration of predefined language specific document feature reduction methods including tokenization, normalization, stop word removal and the other feature selection methods including statistical feature list, gazetteer feature list and regular expiration patterns for some specific feature list. The AOIES prototype consists of four basic phases listed as document browsing phase, preprocessing phase, machine learning phase which is containing two submodules as text classification and information extraction and lastly the postprocessing phase. In this study, the text classification module is trained and evaluated with 1500 sport news instances as a dataset and the information extraction model is trained and evaluated with 1800 sport news instances as a dataset. While evaluating the performance of developed system prototype, the standard precision, recall and F-score evaluation metrics and 10-fold cross-validation method are applied in this study. The experimentation result shows the text classification module achieved 91.7% F-score and the overall IE model has been achieved 94.6% F-score performance by correctly predicting the candidate instances and assigning to their respective category lists. The integration of statistical, gazetteer and regular expression features paly an essential contribution to generate the Naïve Bayes classification algorithm with good performance achievement in this study. From the evaluation result, we can see that the developed system prototype has achieved favorable good performance while extracting relevant football news information from unstructured Afaan Oromoo sport news document. Generally, this study proves that the Naïve Bayes classification algorithm can be used as information extraction method and this indicate it is possible to adopt the supervised machine learning approach for Afaan Oromoo text documents with trivial modification and some linguistic information integration.

References

[1] Stephen Marsland, Nitin Indurkhya, Fred J. Damerau Ralf Herbrich and Tore Graepel, HANDBOOK OF NATURAL LANGUAGE PROCESSING, New York: CRC Press, 2014.

[2] Ronen Feldman and James Sanger, THE TEXT MINING HANDBOOK, New York: Cambridge University Press, 2007.

[3] Hércules Antonio do Prado and Edilson Ferneda, Emerging Technologies of Text Mining: Techniques and Applications, Brazil: InformatIon scIence reference, Hershey • New York, 2013.

[4] Gon»calo Sim~oes, Helena Galhardas, Lu¶³sa Coheur, "Information Extraction tasks," 2013.

[5] h. Extraction, "Information Extraction Tasks and Evaluation," Wekipedia, 2016.

[6] Jie Tang, Mingcai Hong, Duo Zhang, Bangyong Liang, and Juanzi Li , "Information Extraction: Methodologies and," 2014.

[7] G. D. DINEGDE, Afaan Oromoo news text summarizer, 2012.

[8] wikipedia.org, "wikipedia.org," 2014. [Online]. Available: https://en.wikipedia.org/wiki/Information_extraction.

[9] J. P. a. R. Yangarber, Information Extraction: Past, Present, 2014.

[10] Kuspriyanto,Oerip S Santoso, Dwi H Widyantoro,Husni S Sastramihardja,Siti Maimunah, "Performance Evaluation of SVM-Based Information," *Kuspriyanto,Oerip S Santoso, Dwi H Widyantoro,Husni S Sastramihardja,Siti Maimunah*, Vols. Volume 2,, 2013.

[11] wikipedia.org, "wikipedia.org," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Precision_and_recall.

[12] JORDI TURMO, ALICIA AGENO, NEUS CATAL, "Adaptive Information Extraction," Spain, 2014.

[13] Muawia Abdelmagid, Ali Ahmed and Mubarak Himmat, "INFORMATION EXTRACTION METHODS AND EXTRACTION TECHNIQUES IN THE CHEMICAL DOCUMENT'S CONTENTS: SURVEY," *ARPN Journal of Engineering and Applied Sciences* , Vols. VOL. 10, NO. 3, FEBRUARY 2015 , no. ISSN 1819-6608 , 2015.

[14] A. S. Genemo, AFAAN OROMOO CANDIDATE FEATURERECOGNITION USING, 2015.

[15] G. Tsedalu, INFORMATION EXTRACTION MODEL FROM AMHARIC, 2010.

[16] "Alberto Téllez-Valero, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda," *A Machine Learning Approach to Information Extraction* , 2015.

[17] Alberto Téllez-Valero, Alberto Téllez-Valero and Alberto Téllez-Valero, "Using Machin Learning for Extracting Information From Natural Disaster News Report," *National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico*, 2013.

[18] B. W. Agajyelew, "Information Extraction from Amharic language Text: Knowledge-poor Approach," 2015.

[19] BYLINE, "INFORMATION EXTRACTION," 2013.

[20] M. Ipalakova, "INFORMATION EXTRACTION," UNIVERSITY OF MANCHESTER, London, 2010.

[21] Katharina Kaiser and Silvia Miksch, "Information Extraction: Survey," Vienna University of Technology, Vienna, 2015.

[22] S. Sarawagi, "Information Extraction," *Foundations and Trends in Databases*, Vols. Vol. 1, No. 3 (2007) 261–377, 2016.

[23] S. HIRPASSA, DESIGNING AN INFORMATION EXTRACTION SYSTEM FOR AMHARIC VACANCY ANNOUNCEMENT TEXT, Addis Ababa: ADDIS ABABA UNIVERSITY, 2013.

[24] Douglas E. Appelt and David J. Israel, "Introduction to Information Extraction Technology," *IJCAI-99*, 2014.

[25] Douglas E. Appelt and David J. Israel, "Introduction to Information Extraction Technology," *A Tutorial Prepared for IJCAI-99*, 2014.

[26]

[27] R. Grishman, "Information Extraction: Capabilities and Challenges," in *2012 International Winter School in Language and Speech Technologies*, Tarragona, Spain, 2012.

[28] Leonardo-Rigutini, Automatic-Text-Processing: Machine-Learning Techniques, Research Gate, 2010.

[29] M. L. Kejela, "Candidate featureRecognition for Afan Oromoo," *Candidate featureRecognition for Afan Oromoo, AAU*, 2010.

[30] M.-H. ABUBEKER, PART OF SPEECH TAGGER FOR AFAAN OROMOO LANGUAGE, 2010.

[31] G. G. Eggi, Afaan Oromoo Text Retrieval System, ADDIS ABABA, ADDIS ABABA

UNIVERSITY: ADDIS ABABA UNIVERSITY, 2012.

[32] A. D. GEMECHU, "AUTOMATIC CLASSIFICATION OF AFAAN OROMOO NEWS TEXT," ADDIS ABABA UNIVERSITY, DEPARTEMENT OF INFORMATION SCIENCE , ADDIS ABABA, 2009.

[33] F. B. Tesema, Afaan Oromoo Automatic News Text Summarizer Based on, Addis Ababa University, Addis Ababa: Addis Ababa University, 2013.

[34] A. A. Diro, AUTOMATIC MORPHOLOGICAL SYNTHESIZER, ADDIS ABABA: ADDIS ABABA UNIVERSITY, 2010.

[35] João Cordeiro, Pavel Brazdil, "Learning Text Extraction Rules, Without Ignoring Stop Words," *PRIS,* Vols. 128-138 , 2004.

[36] w.Chambling, "Text Classification," 2012.

[37] R. Grishman, Information Extraction:, Tarragona, Spain: Rovira i Virgili University, 2012.

[38] T. G. Debela, AFAAN OROMOO SEARCH ENGINE, Addis Ababa: Addis Ababa University, 2010.

[39] C. F. Elanso, Afaan Oromoo List, Definition and Description, Addis Ababa: Addis Ababa University, 2016.

[40] G. Tsedalu, "INFORMATION EXTRACTION MODEL FROM AMHARIC," *A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE ,* 2010.

[41] https://en.wikipedia.org/wiki/, "Precision_and_recall," 205.

[42] Yugal kumar,G. Sahoo , "Analysis of Parametric & Non Parametric," *I.J. Information Technology and Computer Science,,* 2012.

[43] Shivani Sharma,Saurabh Kr. Srivastava , "Review on Text Mining Algorithms," *International Journal of Computer Applications (0975 – 8887),* vol. Volume 134, no. No.8, 2016.

[44] Vaibhav C.Gandhi, Jignesh A.Prajapati , "Review on Comparison between Text," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) ,* vol. Volume 1, no. Issue 3, 2012.

[45] C. KOIRALA, "COMPARISON OF THE EFFECTS OF LEXICAL AND ONTOLOGICAL INFORMATION," 2013.