[1]Manisha Deka

[2] Prof. Shikhar Kr. Sarma

Identification of Assamese Question sentence using SVM and Naïve Bayes

Abstract: -This paper presents an approach to identify Assamese question sentence from an corpus. It is an important step towards question sentence detection from a document. In this paper, the question structure of different question type in Assamese language is analyzed. We deliberated the lexical feature, syntactic feature and semantic features of question. Syntactic and lexical feature is used to identify Assamese question. We have used SVM, Naïve Bayes for categorization Assamese sentence using syntactic and lexical pattern. We have created our own dataset for training and testing and get maximum of 92% accuracy.

Keywords: Assamese Question Pattern, Question sentence, Feature, Machine learning

I. INTRODUCTION

In online communication, the internet user talks about what they think and feel about the thing in everyday life. This data is not structured. Sentence similarity can be used in text mining as a criterion to find hidden knowledge from textual collections. A question answering system can be more automated using automatic question identification which can examine the interaction between clients and human support staff over a chat channel. The question analysis and categorization are the first steps in the question answering system that enhance the standard of answer. Different syntactic, semantic and lexical rules are used to identify the question phrases effectively and answer quickly.

The structure of question is very important for question answering system to handle the different types of question. The information retrieved from question is required to generate answer. In search engine, results are often far from adequate due to a complex query. Formulation, categorization and entity extraction are the main parts of query evaluation. The question classifier utilizes machine learning with labeled question to identify text chunks by per-established classification. It uses various features like syntax, semantics and morphology [1]. There are some simple approach like pattern matching of each question type for categorization question, deploying the question word position and the different question word. Deep learning approach gives very good accuracy for categorization. The limitation of the resource of language arises problem in deep learning. The accuracy and speed can be increased using neural network as compared to traditional approach.

Past researches on question answering are designed for constrained area and the system performs very limited functionality. But the recent question answering system concentrate clearly on the types of queries submitted by users, the dataset used and the format of the proper replies generated [2]. In recent research, the author uses word embedding to encode text with CNN for text classification.

Question identification is a system of identifying question sentence from a collection of text. It is an influential part for analyzing question sentence structure. The pattern of question in a language is very important for automatic question generation. Question sentence identification is a problem in NLP to extract pattern of question and put forward to mitigate automatic question generation system. Formal text is used to train traditional natural language processing tool lead to difficulty for inspecting question that is searched. The shortage of uniformity in question patterns arises problem to train the model that can analyze question to bring out information. Recognizing well-proportioned natural language question can expedite a rustic interconnection between a user and a machine in chatbots.

In this paper, we have studied the different types of question present in Assamese language and their pattern. We study the lexical and syntactical feature of Assamese sentence. We use two machine learning approaches

[1]*Research scholar, Department of Information Technology, Gauhati University, Assam, India

[2]Professor, Department of Information Technology, Gauhati University

[1]. manisha.deka13@gmail.com [2] sks001@gmail.com

SVM and NB and train the model using lexical and syntactic feature extracted from Assamese sentence. Finally, we design a model to predict a sentence from text into a question and non-question in Assamese language.

II. RELATED WORK

A significant number of researches are done for natural language query in many languages. The research are characterized by question classification, automatic question generation, question taxonomies, question feature, duplicate question detection, well-formed natural language question identification, question-pattern revising in question answering system.

The author in research work [3] classifies the query in two phases using SVM, single valued decomposition (SVD) and Latent semantic index. A dataset of size 10000 with 10 classes and a set of queries in different fields have been tested. The method acquired 98% classification accuracy. An efficient question pattern revising method was introduced in [4]. The system extracts some question pattern and then tries to revise this pattern to reduce the error propagation problem. The procedure of the system is strengthened by using Multi-level encoding and multi-dimensional information. A categorization approach on OE pipeline uses pattern learning approach to learn the pattern of query from question templates and expound the question to convert them into queries [5]. An n-gram model has developed for Bangla sentence structure verification. The structure validity of test sentence are recognized at a rate of 93% [6]. The author identifies the natural language question which is structurally sound and constructs a question dataset of 25,100 which is publicly accessible. The system classify question into two category one is well-formed and other is non-well formed category and gives 70.7%accuracy on test set [7]. A feed-forward neural network with 2-hidden layers with Relu activation on each layer is used. Character-3, 4-grams and word-1, 2-grams, syntactic feature are extracted. A research work on question paraphrase identification was presented using character n-gram embedding [8]. The system uses Quora Corpus and neural architecture with pretrained word embedding. An automatic query reformulation system using reinforcement learning framework was designed to maximize the number of relevant documents [9]. The author used RNN on candidate terms to select one term at a time until the system founds special candidate. A system to detect duplicate question is designed for measuring whether a specified set of question has similar meaning or not [10]. Word embedding, semantic similarities, Siamese Neural Networks are used and get good result in identifying duplicate question.

From our literature survey, we know that accurate search needs a major challenge in trading with vast information. Question answering offers a solution categorizing queries in natural language for efficient responses. The quality and successfulness of question answering increases with efficient research in question classification. There are lots of works in question classification, preprocessing noise but little work on question detection from text or review. So, we take initial steps to classify sentences of Assamese language into question and non-question. Our key attention is to categorize sentence from a dataset to two classes that is question class or non question class.

III. TYPE OF ASSAMESE QUESTION

In Assamese language, question word can be present in different position of question sentence. Most question answering system focus on factoid question. Factoid question are factual in nature and they give single answer. In English, generally factoid question start with Wh-word. But in Assamese language, question word is not always in first position.

In Assamese language, a simple sentence may be of five different kinds. Question can be posed on each Assamese sentence type. From the different types of Assamese sentences, seven types of Assamese question can be formed.

A. One word question sentence

This type of sentence contains only a verb or the complete sentence can be entirely represented by the verb. For eg-

খামনে?
গলনে?

The common form of a single word question is

Main verb + Question word

B. Equational sentence type question

In this type, question can be formed from equational sentence. Different form of question sentence can be referred based on this type of question. For eg:

জিমী কেনেকুৱা?
তাই কোন?
কোন ভাল?
কি ৰঙা?

From the first two questions, a common form can be generated for question as
Noun or pronoun + Question word
The general structure of the question for the last three questions is

Question word+ Noun or Adjective

C. SV or SOV sentence type question

Question can be posed for SV or SOV type of sentence individually to search distinct information. We take some examples of following questions

কোনে খালে?
কি পালে?
কোনে মাৰিলে?
কিহেৰে ৰামে স্কুল গ'ল?

The general form of the above question is
Question word+ Noun + verb
For the another set of questions

গীতাই কি আনিলে?
সি কাক মাতিছে?
জুবিনে কি গান গালে?

The general structure of above question is Noun+ Question word+Noun+verb

Again for this type of question like গীতা কলেজলৈ কি দি গ'ল?
one more format can be derived as
Noun+Question word+verb

D. SOV or OV sentence type question

In SOV or OV type of sentence, there may be two occurrences of verbs. The first verb is the non-finite verb and the other is the finite verb. The question of this type of sentence is like

কোনে পঢ়িবলৈ গ'ল?
কোনে কলেজলৈ গ'ল?
কোনে মাছ ধৰি বজাৰলৈ গ'ল?

The general form of this type of question is
Question word+ Noun+Auxiliary verb +Noun+Verb

E. SVV (non finite) OV (finite) type question

In this type, the question can be asked in distinct ways and separately as

কাৰ ভাগৰ লাগিল?
একেখন আলোচনী সদায় পঢ়ি পঢ়ি কাৰ বেজাৰ লাগিছে?
কোনে পঢ়িলে?

The general pattern are derived for this type of question are

1. Quantifier + Noun+ Manner+ Auxiliary verb+ Question word+ Quantifier +Noun + Manner +Auxiliary verb+Adverb+ Main verb
2. Question word+ Main verb

F. Complex sentence type question

Complex sentence are formed by combining the main clause and the subordinate clause. A set of complex sentence Questions are

ৰামে আবেলি কি দি কলেজৰ পৰা আহিব?
কোনে দুপৰীয়া গা ধুই ভাত খায়?
কেনেকুৱা বস্ত খালে শকত হয়?

The general pattern for this type of question are

1. Noun+Auxiliary Verb + Question word+ Adverb+Auxiliary verb +Noun+Adverb + Main verb
2. Noun+Auxiliary verb +Question word+Noun+Auxiliary verb +Noun + Main verb
3. Question word + Noun + Adverb +Auxiliary verb +Adverb +Main verb

G. Compound sentence type question

In Compound sentence, more than one clause are combined. A set of Compound sentence type question are

কোন এজন ভদ্র আৰু উদাৰ ব্যক্তি আছিল?
কোন চোকা ছাত্র আছিল?
কাৰ বাবে ৰামৰ আকনো মৰম নাছিল?
ৰাম এজন কেনেকুৱা লৰা আছিল?

The general structure of this type of question are

1. Noun+ Question word +particle+Noun +Modifier+Noun + Main verb
2. Noun+ Modifier+ Question word +Noun + Main verb

IV. FEATURES OF ASSAMESE SENTENCES

For Assamese sentence, the selection of best set of feature is difficult for machine learning based approach. In a research work [11], the author studied feature extraction from text for classification. A sentence Q with m word can be taken as

$Q= W_1 W_2……...W_m$

Here W_k is a word in a sentence where 1<=k<=m. From the word of sentences we can choose various type of feature for categorization purpose. These features can be grouped into three categories [12]. The three different types of features are lexical feature, syntactic features and semantic features.

A. Lexical features

The context word present in a question is lexical feature of the question. In the research work [13], the author discuss about the five features. The author used Wh-word, Wh- word position, Wh-type , question length and End marker as a lexical feature.

Our dataset contain the different Wh-Word of Assamese question. The table lists the Wh_word of Assamese language.

Table1. Assamese Wh_word

কি	কিয়	কি কি
কোন	কিমান	কোন কোন
ক’ত	কেনে	কাৰ
কেতিয়া	ক’লৈ	ক’ৰ

In paper [4], it is specified that a question can be presented as the same as document presentation using vector space model.

$q=(q_1, q_2 \ldots . q_n)$

Here q_i is the frequency of i^{th} term in question q and n is the total number of terms.

A question can be regarded as $q=\{(t_1, f_1), (t_2,f_2),\ldots...(t_p,f_p)\}$ where f_i determines the frequency of i^{th} term of a question. In our system, we use Unigram as a lexical feature for identifying question sentence.

Lexical feature of Assamese question “ ছ’চিয়েল নেটৱর্কিঙৰ শক্তিশালী মাধ্যমসমূহ কি” can be listed as

Table2. Lexical feature example of Assamese Question

Feature Space	Feature extracted
Unigram	{(ছ’চিয়েল,1)(নেটৱর্কিঙৰ,1)(শক্তিশালী,1) (মাধ্যমসমূহ,1)(কি,1)}
Bigram	{(ছ’চিয়েল-নেটৱর্কিঙৰ,1)(নেটৱর্কিঙ শক্তিশালী ,1)(শক্তিশালী-মাধ্যমসমূহ,1)(মাধ্যমসমূহ- কি,1)}
Trigram	{(ছ’চিয়েল-নেটৱর্কিঙৰ-শক্তিশালী,1)(নেটৱর্কিঙৰ-শক্তিশালী-মাধ্যমসমূহ,1)(শক্তিশালী-মাধ্যমসমূহ-কি,1)}
Wh-Word	{(কি,1)}

B. Syntactic features

The syntactical structure of a question determines the syntactic feature. A number of syntactical features are derived in different research work and used with different approaches. The two most common kinds of syntactic feature used in the different research work in [13][14][15] for sentence classification are part of speech (PoS) tags or tagged Unigram and head words. PoS determines the part of speech tags of eachword in a sentence like N_NN(Common noun), N_NNP(proper noun), P_PRP(Personal pronoun), P_PRQ(Wh-word), V_VM(Main verb),V_VAUX(Auxiliary verb) etc.

The main word in a question sentence is considered as a head word. It is the word for which question seeks. For determining main word of a sentence, we need a syntactic parser. But there is no efficient syntactic parser for Assamese language. In our work, we identify question sentence from a corpus of Assamese assertive and question sentence. Since we consider assertive sentence, PoS tag is used as syntactic feature.

C. Semantic features

The relationship between different words can be determined by semantic feature. The semantic meaning of a word in a query can derive the semantic feature. Wordnet is a lexical database which can provide higher level of semantic [16]. Wordnet is used to find out feature. In our system, semantic feature is not used as we do not have any efficient resource for determining this feature.

V. MACHINE LEARNING APPROACHES FOR QUESTION IDENTIFICATION MODULE

Different machine learning approaches have been used for different classification problem. We use machine learning classifier in our system for Assamese question sentence identification. We work with two classes for Assamese sentence. One class is assigned for Assamese question sentence and another class for Assamese assertive sentence. The aim of our question identification module is to set class label to a Assamese sentence. We need a feature matrix and a fit algorithm to design machine learning based system. The problem, data size and feature set determine the fulfillment of an algorithm. To do this, we have applied SVM, Naive Bayes approaches for identification of Assamese question from a corpus.

A. SVM

SVM is a classifier that separates instances with a hyper line. It makes an optimal straight line which generates categories. It can operate in fairly large volume of feature set. Classified documents are used for training SVM model [17].The SVM algorithm perceives a line with largest margins between categories as the ideal separating hyperplane maximizes the margin of the training data.

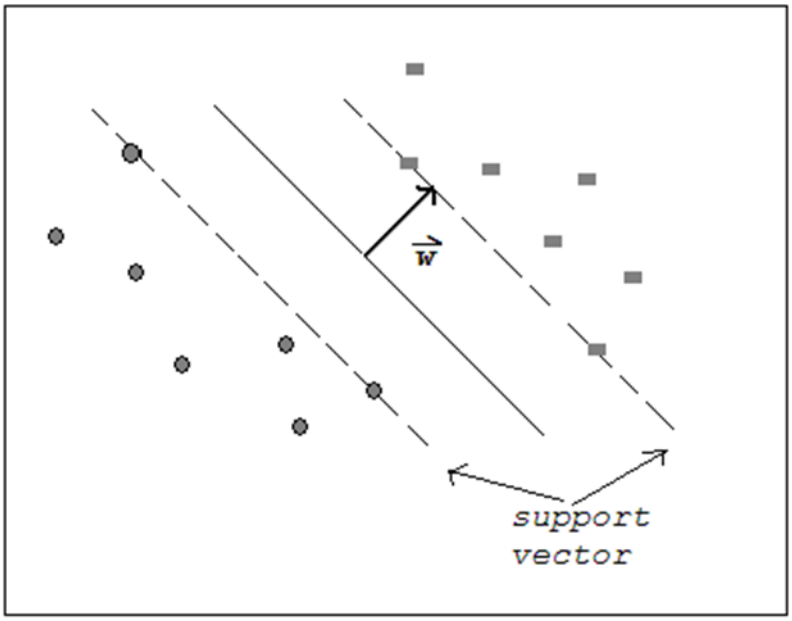

Fig. 1. Example of SVM hyperplane pattern

SVM is a powerful statistical learning technique to solve different kinds of problems [18]. It is suitable to manage complex data like text and image with high dimensions [19]. The SVM algorithm gives efficient result in pattern recognition domain [20].

we employ SVM in our paper as this algorithm set up hyper plane to classify the training data with high speed and more accuracy compared to other traditional method

B. Naive Bayes

Naive Bayes classifier is a straightforward probabilistic classifier placed on applying Bayes' theorem with spowerful independence premise. The Naive Bayes Algorithm is known for its simpleness and efficiency. Using this approach, model building and prediction are completed more quickly. When previous knowledge is available, the Bayes theorem is used to calculate the probability of a hypothesis. Conditional probabilities are the basis for it.

The formula for conditional probabilities is

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is the probability of a hypothesis A, given that event B occurs. P(B|A) is the probability of the evidence given that hypothesis A is true. P(A) is the probability of the hypothesis before observing the evidence, and P(B) is the probability of the evidence.

Naive Bayes is used to solve different problem in NLP research. To classify quiz question, this algorithm was used and achieved 91% accuracy [21]. In Research work [22], numeric and location type question are classified using this algorithm. This algorithm is used in different research work for sentence categorization and get very good result [23][24][25][26].

VI. Implementation and result

A. Corpus preparation

Assamese is a low resource language. There is no standard question data available till date. This limitation of the resource in Assamese creates problem in deep analysis for question sentence. To solve this problem, we have created a question dataset for experiment. We have collected some document from different domain. A total of 2000 question are prepared and annotated with PoS tag. In our previous work, we have found 517 unique PoS pattern from this question dataset. We again prepared some new question and add to the previous dataset. This question sentence is mixed with some assertive sentence to train our system so that it can classify question sentences and assertive sentence. Our dataset contain 5000 sentence. Each data point in the dataset is a question sentence or an assertive sentence.

B. Implementation

Our goal is to analyze Assamese question sentence and to categorized Assamese sentence into two different class. One class is Assamese assertive sentence and other is Assamese question sentence. For this, we use a corpus of Assamese sentence. The corpus was tagged using Assamese PoS-taggers that have been already developed using BIS-tagset. We have used two models Naive Bays and SVM for categorization of Assamese sentences. For classification, Lexical and syntactic feature are used. SVM and NB model are trained and tested using NLTK. The whole dataset is divided into 80% to train and 20% to test the model. TF-IDF is used for converting the training and testing the data into numerical feature vector. After this SVM and NB model are trained and tested. The proposed model will take an Assamese sentence and can identify whether it is question sentence or assertive sentence.

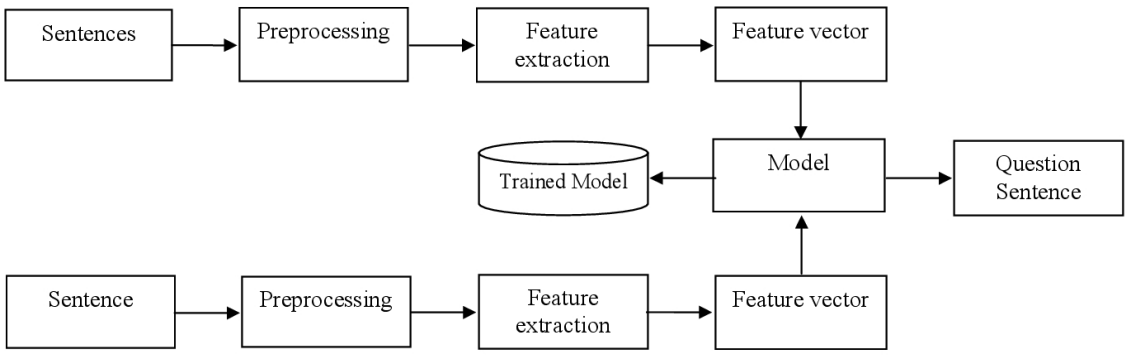

Fig. 2. Architecture for question sentence identification module

C. Result Analysis

We have conducted experiment on a training set of 4000 question. The model is tested on 1000 Assamese sentences to identify Assamese question sentence. The result of the experiment is evaluated using classification accuracy. In our experiment, we get good result for SVM as compared to NB. In case of SVM,we get 92% accuracy using syntactic feature where as NB gets 85% accuracy. Similarly, using lexical feature 85% accuracy is achieved for SVM and 80% for NB. SVM gives better result compared to NB on both feature.

Table 3. Accuracy of NB and SVM with different training set size using Lexical feature and Syntactic feature

Training data set size	Lexical feature		Syntactic feature	
	NB	SVM	NB	SVM
1000	65	68	69	72
2000	70	72	73	79
3000	76	81.5	81	88
4000	80	85	85	92

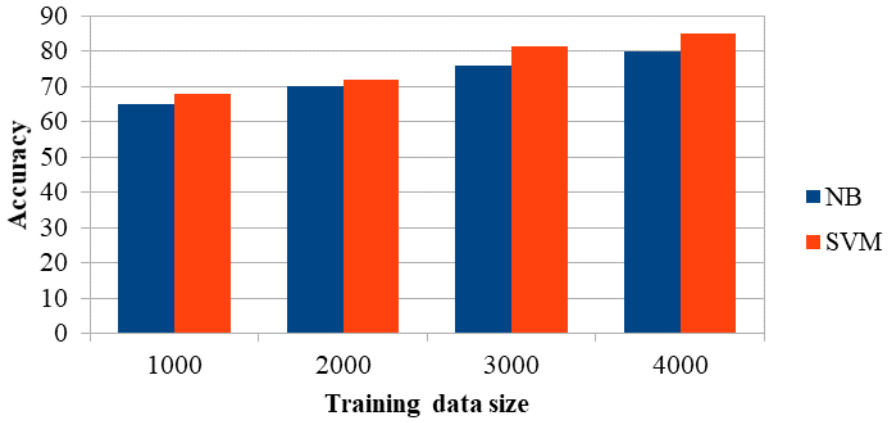

Fig. 3. Accuracy Vs training data size with lexical feature

Table 4. Performance evaluation of NB and SVM in terms of Accuracy

Model	Syntactic feature	Lexical feature
SVM	92%	85%
NB	85%	80%

VI. CONCLUSION

In our work, we proposed a new task of natural language question sentence identification for Assamese language. First, we analysis the structure of the different type of Assamese question. We generate a dataset of question sentence and assertive sentence of Assamese language. Then we extracted the lexical and syntactical feature of these sentences. We design our model using SVM and NB and train model to predict the class of a sentence whether it is question or non question. Our work is the first attempt in Assamese language to classify Assamese sentence. Our system achieved accuracy of 80% for NB and 85% for SVM using lexical feature and 85% for NB and 92% for SVM using syntactic feature on test set.

Question formedness information is useful for improving state-of-the art automatic question generation system. Our work can be extended to identify a well-formed question in Assamese language which can enhance query understanding. We can increase our system performance by using deep leaning method in future.

References

[1] Nadir Hussain, Sheikh Muhammad Saqib, and Muhammad Usman Gurmani, “Detection of questions from Text Data Using LSTM-Deep Learning Model,” VAWKUM Transactions on Computer Sciences , March 2024

[2] A. Mishra and S. K. Jain, “A survey on question answering systems with classification,” vol. 28, no. 3, pp. 345–361, Jul. 2016, doi: 10.1016/j.jksuci.2014.10.007

[3] Moayeah Al_shenak, Khalid M. O. Nahar, Khaldoun MK. H Halawani “AQAS: Arabic Question Answering System based on SVM, SVD and LSI,” Journal of theoretical and Applied Information Technology, Vol. 97, No 2, 2019

[3] Yanchao Hao, Hao Liu, Shizhu He, Kang Liu, and Jun Zhao, "Pattern-revising Enhanced Simple Question Answering over Knowledge Bases," Proceedings of the 27th International Conference on Computational Linguistics, pages 3272–3282 Santa Fe, USA, August 20-26, 2018

[4] Hapnes Toba and Mirna Adriani, "Pattern Based Indonesian Question Answering System," https://www.researchgate.net/publication/258924979

[5] Nur Hossain Khan, Md. FarukuZZaman Khan, Md. Mojahidul Islam, Md. Habibur Rahman, and Bappa Sarker, "Verification of Bangla Sentence structure using N-Gram," Global journal of computer science and technology: A hardware &computation, Volume 14, Issue 1, 2014

[6] Manaal Faruqui and Dipanjan Das, "Identifying Well-formed Natural Language Questions," Proceedings of the 2018 Conferrence on Empirical Methods in Natural Language Processing

[7] Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das, "Neural Paraphrase Identification of Questions with Noisy Pretraining," Proceedings of the First Workshop on Subword and Character Level Models in NLP, pages 142–147, 2017

[8] Rodrigo Nogueira and Kyunghyun Cho, "Task-Oriented Query Reformulation with Reinforcement Learning," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[9] Kumar Saksham, Dr. Pawan Kumar, and Dr. Mir Aadil, "Duplicate Question Pairs detection using NLP,"International Journal of Advanced Research in Computer and Communication Engineering, 2023

[10] Abdelwadood Moh'd A MESLEH, "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System" Journal of Computer Science, Volume 3, No. 6, 2007, 430-435

[11] Babok Loni, "A Survey of State-of-the-Art Methods on Question Classification," TU Delft Research Repository 2011

[12] Somnath Banerjee and Sivaji Bandyopadhyay, "Bengali Question classification: Towards Developing QA System," Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, 2012

[13] Xin Li, Dan Roth, and Kevin Smal, "The Role of Semantic Information in Learning Question Classifiers," International conference on computational Linguistics

[14] Phil Blunsom, Krystle Kocik, and James R. Curran, "Question classification with log-linear models," SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Pages 615–616, August, 2006

[15] Nguyen Van-tu and Le Anh-Cuong, "Improving question classification by feature extraction and selection," Indian journal of science and Technology, Vo9(17)

[16] Jin Huang, Jingjing Lu, and Charles X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy," Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), 2003

[17] A. Basu, C. Watters, and M. Shepherd, "Support Vector Machines for Text Categorization," Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03) ,IEEE, 2003

[18] Vladimir N. Vapnik, "The nature of statistical learning theory " Springer science & business media, 2013

[19] Y. Benayed, D. Fohr, J. P. Haton, and G. Chollet, "Confidence measures for keyword spotting using suport vector machines," IEEE International Conference on Acouustics, Speech and Signal Processing, vol 1, 2003

[20] Annisa Syafarani Callista,Oktariani Nurul Pratiwi, and Edi Sutoyo, "Questions Classification Based on Revised Bloom's Taxonomy Cognitive Level using Naive Bayes and Support Vector Machine," 4th International Conference of Computer and Informatics Engineering (IC2IE), IEEE, 2021

[21] Jeena Mathew and Shine N Das, "Question Classification using Naive Bayes Classifier and Creating Missing Classes using Semantic Similarity in Question Answering System," International Journal of Engineering Trends and Technology (IJETT), Vol 23, 2015

[22] Arun D Panicker, Athira U, and Sreesha Venkitakrishnan, "Question Classification using Machine Learning Approaches," International Journal of Computer Applications Vol48, No.13, June 2012

[23] Novi Yusliani, Syechky Al Qodrin Aruda, Mastura Diana Marieska, Danny Mathew Saputra, and Abdiansah Ns, "The effect of Chi-Square Feature Selection on Question Classification using Multinomial Naïve Bayes," Sinkron : Jurnal dan Penelitian Teknik Informatika" Volume 7, Number 4, October 2022