

Cover sheet for submission of work for assessment



UNIT DETAILS

Unit name Technology in an Indigenous Context Project Class day/time Tuesday (8:00 - 12:00) Office use only

Unit code COS10025 Assignment no. 2 Due date 2/7/23

Name of lecturer/teacher Dr. Nguyen Phuong Anh

Tutor/marker's name Dr. Nguyen Phuong Anh Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Nguyen	Ha Huy Hoang	103487444
(2)	Ta	Quang Tung	104222196
(3)	Vu	Minh Quang	104184089
(4)	Do	Tuan Dat	103804603
(5)	Tran	Khai Kiet	104085315
(6)			

DECLARATION AND STATEMENT OF AUTHORSHIP

- I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
- This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
- No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
- I/we have not previously submitted this work for this or any other course/unit.
- I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.

I/we understand that:

- Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.

Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

(1)	Nguyen Ha Huy Hoang	(4)	Do Tuan Dat
(2)	Ta Quang Tung	(5)	Tran Khai Kiet
(3)	Vu Minh Quang	(6)	

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at www.swin.edu.au/student/

Copies of this form can be downloaded from the Student Forms web page at www.swinburne.edu.au/studentforms/

Teamwork breakdown

Family Name	Given Name	Contribution	Words
Nguyen	Ha Huy Hoang	- Part A project overview, requirements - Part B design concept 2	1539
Ta	Quang Tung	- Part A project overview, requirements - Part B design concept 4	1533
Vu	Minh Quang	- Part A project overview, requirements - Part B design concept 1	1255
Do	Tuan Dat	- Part A project overview, requirements - Part B design concept 3	1386
Tran	Khai Kiet	- Part A project overview, requirements - Part B design concept 5	1302

Table of Contents

Part A	3
Project overview (everyone)	3
Requirements	4
Part B	5
I) Design concept 1: Personalize training with recommendation system	5
1) Learning issues	5
2) Description	5
II) Design concept 2: A blockchain system to connect business and farmers	8
1) Learning issues	8
2) Description	8
III) Design concept 3: Search engine with NLP voice recognition to get accurate search ...	12
1) Learning issue	12
2) Description	12
IV) Design concept 4: Misinformation checking with Deep learning	15
1) Learning issue	15
2) Description	15
V) Design concept 5: Auto translation with Transformer model.....	19
1) Learning issue	19
2) Description	20
References	24

This report is divided into two parts. The first part (part A) will give an overview of the project learning issues and outline the requirements of the project. The second part (part B) will go into detail about how the system will deal with the learning issues with specific technologies.

Part A

Project overview (everyone)

Although the Vietnamese government has already implemented many agricultural training programs to help ethnic minorities catch up with current development, most of these programs are not returning good results. According to Nguyen Tung Phong [\[1\]](#), only 43.5% of the training knowledge is applied to reality, which suggests that the current vocational training approach is inefficient and impractical. This project aims to solve this problem by developing a specialized information system for ethnic agricultural training. After extensive research, our team has identified five issues with the current training approach, and we will explain how each of these issues can be addressed with an information system.

First, the existing vocational training approach is not addressing the needs of mountainous residents. According to Ha Thi Hai Do [\[2\]](#), most ethnic workers believe that vocational training is time-consuming. Only 14% of the total ethnic minority population are taking part in these training programs and they are more likely to take short-term classes instead of a 3-month course with full training content. However, with an information system, a personalized training model can be deployed to provide focused training for each individual and bring a much closer program to the need of the mountainous farmers.

Second, the current training is not only out of the interest of mountainous residents but also lacks connection with enterprises. According to Tran Quoc Nhan [\[3\]](#), most of the farmers and private sectors are working with each other through contracts, but there are many factors such as market prices that might affect the will to fulfill the contract on both sides, thus making the connection less appealing. To solve this problem, an information system can use blockchain technology to help farmers and businesses connect, sign contracts, support farmers to fulfill them, and ensure reliability.

Third, mountainous people with low literacy may find it hard to search for information in an information system. As pointed out by Nisansala Vidanapthirana's research [\[4\]](#), one of the biggest challenges of an information system for underdeveloped rural areas is illiteracy, since television shows are the only source of information viable for people who do not know how to read and write. Therefore, the use of a high-spec information system that supports voice recognition for searching will help these marginalized people find the information they need.

Fourth, the difference in languages among Vietnamese ethnic groups can also be a big hindrance for the ethnic minorities to learn from resources recorded in general Vietnamese. According to

GSO's statistics in 2019 [5], only 80,9% of the ethnic population knows general Vietnamese, which means nearly one out of five people in mountainous areas gets alienated due to language. However, an information system can be a perfect solution to this with the implementation of an auto-translation AI model.

Finally, the last learning issue is the problem of misinformation. During the Covid-19 epidemic, the spread of misinformation caused great harm to the whole society and led to many misconceptions after the quarantine period. Wenjing Pian pointed out that there are 3 main causes of this misinformation epidemic: the use of social media, fast publication, and the low health literacy of Internet users [6]. With the introduction of an information system that allows people to share their knowledge on the platform with ethnic minorities who have low literacy, misinformation can be a serious problem. Therefore, the implementation of an information-checking system will be an important task to protect mountainous farmers.

Requirements

This section will discuss the requirements for the functions listed above. Each requirement will focus on the efficiency of a function and determine the threshold at which it can be considered finished.

The first function is a recommendation system. This function should be able to give advice based not only on the input information from the users but also on auto-connected data such as online weather data or IOT devices. In this way, the system will depend not solely on the input information but also on other more accurate data sources. Moreover, the recommendation system should also have at least 90% accuracy during the testing phase. With this requirement, the recommended tutorial delivered to mountainous residents will bring back good results.

The second function is a blockchain contract system to connect with private sectors. This system should be able to check the input to validate if the specifications of agricultural products meet the requirements of private sectors with a smart contract system powered by machine learning, thus ensuring the safety of the contract. Moreover, the information about these contracts should be kept private from outside hackers and information thieves with encryption mechanisms.

The third function is voice recognition searching. The search system should apply deep learning algorithms to get the voice content from the user and extract its feature before giving them feedback. The returned results should give accurate recognition and also return accurate values based on the recognized speech.

The fourth function is video content translation. As discussed above, the language barrier is one of the biggest problems to deliver quality knowledge content to remote areas. However, with the power of a transformer model, we are now able to break this barrier with highly accurate translation between languages and bring knowledge to the mountainous areas easier. For this reason, accurate and understandable translation is an inevitable requirement of this system.

Finally, the last function is misinformation checking. This function is required to protect ethnic communities from any misunderstanding or confusion since low literacy makes them vulnerable to misinformation. For this reason, the misinformation detection system should be strong enough to check both the video content and the script content of the videos uploaded to the system with at least 80% accuracy.

Part B

In this part, this report will dive into the detail of the design idea. The idea will need to be based on the learning issue and requirements. To address this, each design concept will be presented with the following contents:

- A brief opening with the identified learning issue: restate the learning issue of the idea
- Design outline: the structure of the idea, how the system works
- Specifications: a list of items and materials needed to deploy the above system.
- Design benefits: the benefits that the idea can bring
- Design constraints: the problem that might cause harm to the project.

I) Design concept 1: Personalize training with recommendation system (Vu Minh Quang)

1) Learning issues

In the preceding discussion, the fact that a majority of the ethnic minority believe that agriculture training programs are time-wasting has been proved. Therefore, to combat this situation, a change in mindset will not be enough solve this situation, but it is the current training program to change to become better.

2) Description

a) Design outline

Today, we may combine individualized training courses to reach out to ethnic minorities. Advanced individualized training that takes into account each person's particular needs, objectives, and preferences is known as personalized training. This style of training enables the customization of the training pace and content to meet the unique needs of each trainee, allowing them to progress more quickly and concentrate on their weak areas. Additionally, students feel driven and want to actively practice while studying with personalized goals. By demonstrating the value and usefulness of the abilities they are learning, learners are more motivated and satisfied.

The problem of information overload that prevents timely access to things of interest has been created by the rapid development in the amount of digital information available and the number of

Internet users. This issue has been largely resolved by information retrieval systems like Google, DevilFinder, and Altavista, but prioritization and personalization (where the system maps the accessible content to the user's tastes and preferences) are still lacking [8]. This has made recommendation systems more important than ever. By selecting the most pertinent information from vast quantities of dynamically created material based on the preferences, interests, or observable behaviors of the user, a recommendation system is an information filtering system that addresses the issue of information overload. The recommendation system is capable of predicting whether a particular user will like an item based on the user's profile.

The suggestion system is advantageous to both the user and the service provider. They lessen the expense of searching per transaction. It has also been demonstrated that recommendation systems can enhance decision-making and quality. For instance, recommendation systems increase revenues in an e-commerce setting since they are a successful way to sell more things [8]. The referral system in scientific libraries helps patrons by enabling them to do searches outside of the catalog. Therefore, it is imperative that systems which provide users with relevant and trustworthy recommendations employ accurate and effective recommendation techniques.

b) Specifications

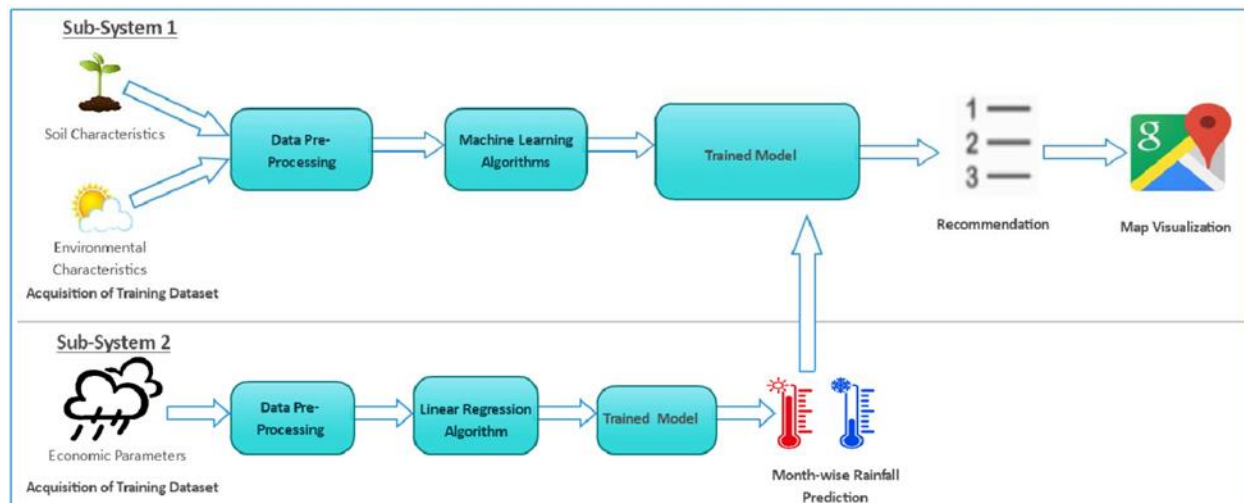


Figure 1. Structure of the recommendation system [7]

A decision-making approach for users in a complicated information environment is referred to as a recommendation system. Additionally, the recommendation system is described from an eCommerce viewpoint as a tool that aids users in searching through knowledge data relevant to their interests and fields. A recommendation system is defined as a way to support and improve the social process of relying on the recommendations of others to make decisions when there is insufficient personal knowledge. The system will choose and present the most pertinent and accurate material when users seek answers or related abilities.

Examples of such software might be: Using collaborative techniques, GroupLens is a news-based architecture that helps users find items in a sizable news database; Amazon utilizes subject diversity algorithms to enhance its suggestions, while Ringo, an online social filtering system, builds user profiles based on ratings for music albums [8].

Any machine learning algorithm's accuracy is influenced by the training dataset's correctness and parameter count. The 'Indian Agricultural and Climate Dataset' was utilized for the first subsystem [7]. The aforementioned metrics are offered for fifteen sub-categories (barley, cotton, peanut, gram, jute, other legumes, potato, ragi), five major categories (bajra, jowar, maize, rice, and wheat), as well as fifteen sub-categories (tur, rapeseed and mustard, sesame, soybean, sugarcane, sunflower, and tobacco).

This involves two steps (Figure 1). The initial step is to eliminate the missing values, which are shown in the original data set as dots. These missing values lower the value of the data, which inhibits the effectiveness of machine learning models. As a result, in order to deal with these missing values, huge negative values are used instead, which the trained model can clearly recognize as outliers. Generating class labels is the second stage before the data is prepared to be applied to the machine learning algorithm. Class labels are required because supervised learning is being used. Since there are no labels in the initial dataset, they must be created at the data preprocessing stage. The yield (in tons) and cultivated area (in hectares) for each crop are used to generate the required labels. A designation of 1 is given to those with a production value of an area larger than 0. A label of type 0 is given in all other circumstances [7].

c) Benefits

The parameters of the farm's environment (temperature, precipitation, latitude, longitude, altitude, and distance from the sea) as well as the properties of the soil will be taken into account by this intelligent system. Helping locals is really convenient, and all of the information is reliable and correct. This technology will let farmers decide wisely which crops to cultivate based on a range of geographical and environmental characteristics.

d) Constraints

People will need time to adjust to and become used to this program because it incorporates numerous tools and technologies. They may find it challenging to change from a direct agricultural method to remote management since their behaviors have been embedded in their traditions and way of thinking for a very long time. This new system will be challenging to be implemented universally if it is not adopted by the young, who are easily acclimated to new technology.

II) Design concept 2: A blockchain system to connect business and farmers (Nguyen Ha Huy Hoang)

1) Learning issues

As discussed above, agriculture is not only about planting crops and raising animals, but it is also about consuming and processing these raw products into high-quality goods. This also means that engagement from both the farmers and the private sectors is needed to form strong and sustainable agriculture. However, this connection seems to be extremely weak in the mountainous areas, where most of the residents work in small-scale household farms. The lack of understanding and skill of the farmers paired with the high logistic cost due to rugged terrain has made these areas less appealing to the private sector.

2) Description

a) Design outline

For the reason above, this project will also become the bridge between the farmers in the mountainous areas and the private sectors by applying the CDIO education model. This model was first introduced by MIT based on 4 criteria: conceive, design, implement and operate to maximize the performance of their students. The students under this training system will first discuss the problems they are facing, then design their idea with the known problem before implementing this idea and testing it. However, with vocational training, this idea will be slightly different. During the conception and design phase, the private sector will use the application to order a specific product and build the requirements of this order (cost and quality) on the application, and the farmers can select an order to follow. After this, the mountainous farmers in the implementation and operation phase will be able to access special training from the private sector and also get the recommendation tutorials from the application to meet the standard [\[11\]](#).

To do this, we will use a private blockchain network to achieve the connection between farmers and businesses. A blockchain system will have the ability to ensure the safety of the transaction, which will act as insurance to protect companies from mistransactions and protect the indigenous community from actions that can harm their reputation [\[12\]](#). This system will also require the farmer to provide information on the specifications of how they raise their products such as pesticide, water, and temperature values.

Many research papers recommended that data fed to the blockchain system should be collected from the farmers through IoT devices such as sensors and cameras to validate the product quality in the smart contract process [\[12\]](#). However, since the target users of this project are indigenous people who are used to working on a smallhold farm, working with technology such as sensors might be costly and ineffective. Therefore, the main method to collect data will require the farmer to input the data by hand.

In this project, our blockchain model will be deployed in a Cloud environment. This blockchain system will also implement a smart contract system to check if the requirement from the enterprises

is addressed with a machine learning algorithm from the farmer inputs. After being analyzed with the AI smart contract system, the qualified products will be listed in a block with their specifications before being sent to the next part of the supply chain system.

With cloud machine learning, user inputs will be thoroughly analyzed and evaluated before the results are sent to the private sector. Farmers will also get advice and recommendation on their products before deploying a full-scale farm. After the products have reached the private standards, the farmers will be able to sign selling contracts with the private sector. These contracts will be put into a block and sent to the company factory.

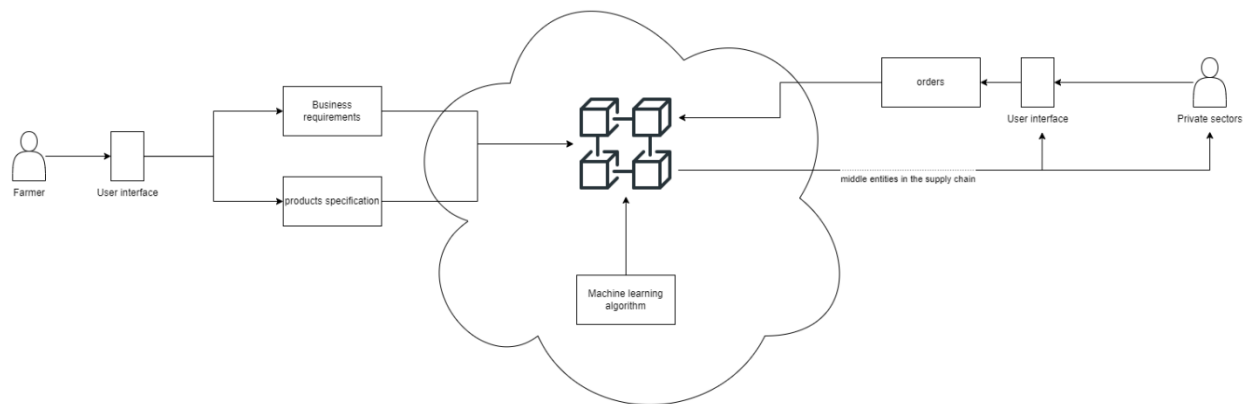


Figure 2. Cloud architecture of the contract system

The picture below is the architecture of a blockchain system used in real estate with deep-learning smart contract verification. The farmers' inputs (which replace the RentData presented in the model) will need to run through a cryptographic transformation (the red zone) to encrypt the data and create an end-to-end password for P2P communication in the Blockchain system [12].

Three components of this transformation will do different jobs. ElGamal Crypto Algorithm will encrypt the input information as well as provide a way to communicate with the encrypted information. Bone-Lynn-Schacham (BLS) will support applying a short signature using the binding properties of an elliptical curve. Finally, Elastic Curve Primitives analysis and Elastic Curve Qu-Vanstone (ECQV) will apply certification-based technology to the system. Smart Contract verification with machine learning will also run parallel with this system for information verification [12].

After the transformation encryption, the data will be pushed to the crypto-implement components before forming a block of data.

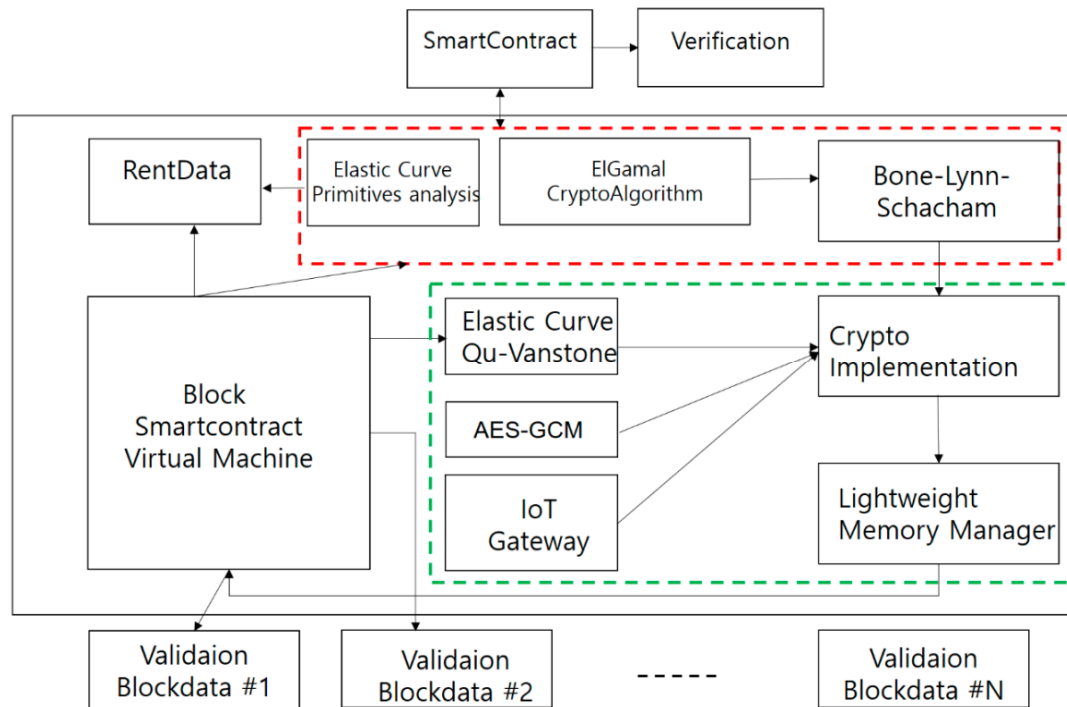


Figure 3. Blockchain system [12]

b) Specifications

To deploy the above blockchain architecture, the use of cloud computing services will be inevitable:

- Container service: the smart contract blockchain line will be deployed onto a container, so a cloud container service will be required to deploy the blockchain infrastructure.
- Computing service: the blockchain and smart contract system will use a virtual machine to run its algorithm so computing service is important.
- Storage service: the use of storage service is nearly unavoidable when creating a cloud architecture so putting it into consideration is also understandable.

With the list of important services above, a comparison between cloud providers will also be needed to get the best cost optimization and performance. Specifically, the comparison below will focus on AWS and Azure cloud services.

Services/providers	AWS	Azure
Storage [13]	S3	Blob
	<ul style="list-style-type: none"> - more functions - costlier but stable cost 	<ul style="list-style-type: none"> - fewer functions - cheaper but costs more money if high-level functions are used
Computing [14]	EC2	Azure Virtual Machine
	<ul style="list-style-type: none"> - Support predictive scaling - Better availability (25 availability zones) - 1 on-demand instance (T4g.xlarge) costs \$101/month 	<ul style="list-style-type: none"> - Does not support predictive scaling - Decent availability (25 availability zones) - 1 on-demand instance (Bs-series) costs \$121/month
Container [15]	ECS	AKS
	<ul style="list-style-type: none"> - Container - Only work for AWS - with spot type and Fargate launch type, ECS only costs \$0.0013335/hour/store 	<ul style="list-style-type: none"> - Kubernetes - work for everything - cost \$0.1/hour/store

Table 1. AWS vs Azure comparison table

As can be seen from the table, AWS is a better option for this project since it has better support for all the services that will be used.

c) Benefits

This implementation can bring many benefits to not only the farmers of the mountainous areas but also help the private sector in countless ways.

For farmers, this application could bring a total transformation to the mountainous area agriculture. The mountain residents will steadily change their working mindset and aim towards a more specific goal instead of working on a smallhold farm without a goal as before. The tutorials and advice from the system as well as from the business experts will help them to build up their skills and understanding about the unique plantation that only their mountain area has to gain an advantage over the lowland area. Moreover, they can also build their reputation and protect it with the blockchain system backtrace function.

For the private sector, this application can provide support in accessing products in mountainous areas. With the use of the blockchain system, the transaction between the private sector and the farmers will be secured, which also means that fraudulent transactions can be minimized. Even if it happens, the private sector can still trace back to find the responsible one with this blockchain system where everyone has their unique identification code. Another benefit of this system is the sustainable ingredient source that it can bring. Private sectors no longer need to find or spend a

long-time training farmers to get their ingredients. Instead, they can find what they want on this system. Moreover, they can also find rare ingredients that can only be obtained in these remote areas just by searching for farmers who have already studied and grown these ingredients and they can also check the quality of the products as well as other specifications about how they are grown.

d) Constraints

Despite bringing many foreseeable benefits, the system still has many weaknesses. One of the obvious disadvantages is the process of how the system obtains data. It will require both farmers and businesses to enter their data into the system which might make it hard to verify if the transaction inputs are true or not. Moreover, the problems with the cost efficiency of this system should also be considered.

III) Design concept 3: Search engine with NLP voice recognition to get accurate search (Do Tuan Dat)

1) Learning issue

Due to the lack of interaction with technologies, it is understandable that most ethnic minorities would have inadequate understanding to react and respond to the elements of the proposed software platform. A demand for supportive features from the platform is established to assist ethnic minorities in utilizing the functions of the software and the information it contains. To cover as many methods as ethnic minorities could use to interact with the system, search engine optimization and natural language processing are integrated into the back-end side of the software, and visibility elements are organized on the front-end side of the software to avoid confusion and enhance readability.

2) Description

a) Design outline

Ease of use is the top priority of the software platform, as most of its users are unfamiliar with digital technologies. Hence, a few criteria for software development are proposed to ensure a smooth user experience. Simplicity will always be promoted to avoid confusion and engage interaction; this includes the use of neutral colors, uncomplicated typography, and the avoidance of unnecessary elements. Consistency must be implemented in the layout of elements on similar pages. This will demonstrate a pattern to the users so that they can intuitively understand the categories of the content contained in different framework blocks. Important information will be highlighted and grouped together to easily draw the users' attention to the crucial content and prevent wasting time in search of needed information.

Another useful concept that we could take advantage of is the idea of integrating objects that ethnic minorities are familiar with into elements of the platform. The representation of these elements

could invoke the operations of familiar objects, giving users a hint of those elements' functions. Continuous engagement with these resources will create links between the functionalities of the platform and the objects that they are familiar with, hence significantly shortening the learning process.

To optimize the search engine within the platform, accurately indexing the content contained in the software is vital; this will prevent confusion when relevant results are not provided in pairs with the correct keywords. In addition, auto-suggestions are applied to the search engine to provide suggested results without the need for complete expressions from the users. This function will be of great assistance to users who may have a hard time expressing their desired content due to low digital literacy. Furthermore, relevant results are displayed in cases where typos occur in the search term. This function will serve users who are not fluent in typing the correct words on a software platform [17].

Natural language processing (NLP) is another important function of the platform. It uses artificial intelligence to process the data that users require. This function can serve as an alternative solution for searching in cases where users are uncomfortable with or unable to type out keywords. The methodology begins with Segmentation where multiple documents are broken down into constituent sentences. Those sentences are then Tokenized, meaning that they get broken down into constituent words to be stored. Non-essential words, called stop words, in the sentences will be removed as they don't add much meaning to the sentences (words such as 'is', 'and', 'an'). Subsequently, Stemming will interpret words with added prefixes and suffixes (words such as 'cries', 'crying', 'cried') to be the same as their version without prefixes and suffixes ('cry'). A similar process, Lemmatization, will identify the base words for different word tenses ('fly' is the present tense of the word 'flew'). Afterwards, Part of Speech tagging will associate parts of the sentences with nouns, verbs, etc. Following this, Named entity tagging will flag names of important entities like locations, weather, or cultural references that may occur in the document. Finally, a machine learning algorithm is used to train the models to return accurate results based on voice inputs [18][20].



Figure 4. NLP model training process

b) Specifications

Several natural language processing services that leverage the required experience in developing machine learning models are proposed in order to take advantage of the technical needs of the machine learning experience, eliminating the need for calculating costs and maintaining data sets. These services will be used to perform natural language processing tasks, providing data to train

the answer-to-question model, resulting in precise responses to the expectations of users' voice inputs.



Criteria	Google Cloud Natural Language [23] 	Amazon Comprehend [24] 
Price	\$0.05 to \$0.2 per 100 characters	From \$0.00005 to \$0.003 per 100 characters
Administration	Easier to administrate	Harder to administrate
Text inspection	Include Part of Speech – tagging, relations in sentences, lemmatization, and morphology analysis	Extract key phrases from the document and return a confidence score
Topic modeling	Discover several chains of category from the text and return a confidence score for each chain	Presents 2 variants. The first variant returns the topic group, main keywords for this group, and confidence score. The second variant returns the topic group and the proportion of main keywords in the document.

Table 2. Google Cloud Natural Language vs Amazon Comprehend comparison table

Google Cloud Natural Language is more expensive but easier to administrate. Amazon Comprehend process text analysis more meticulously than Google Cloud Natural Language, this would make the training process for the model more accurate and effective. However, Amazon Comprehend provides more specific topic modeling, which simplifies keyword searching engine development.

c) Benefits

One of the purposes of the project is to make use of design principles and ethnic minorities' familiarity with their daily life objects to create a user-friendly software environment. Supportive features such as search engine extensions and voice input are also incorporated into the platform to assist them obtain desired results from the software. The ultimate goal of these functions is to

enable a smooth learning and training environment for ethnic minorities so they can benefit as much as possible from the platform.

d) Constraints

There is one constraint in the initial stage of training the natural language processing model. Most of the languages that ethnic minorities use are poorly resourced. Therefore, services like Google Cloud Natural Language and Amazon Comprehend which offer machine learning models don't support these languages. To generate documentation for the natural language processing model to improve accuracy upon returning results for voice inputs, we need to cooperate with linguistic experts of ethnic minority languages to manually translate documents from the source languages to supported languages like English. This will require significant time investment and effort.

IV) Design concept 4: Misinformation checking with Deep learning (Ta Quang Tung)

1) Learning issue

This design concept presents a solution for the detection of misinformation in the system. Misinformation represents a severe risk that must be considered when designing an information system because it can do serious harm to the communities in which it circulates. Assessing the spread of misinformation during the COVID-19 pandemic [5], find that its impacts include raising psychological issues among the public such as anxiety and depression, prompting people to ignore safety guidelines, undermining the public's trust in the government and healthcare system, and causing a host of other social problems. Two factors contributing to this spread are social media platforms that provide a place to share misinformation and the low level of health literacy among the public that causes them to have difficulty absorbing information and making health decisions.

Given that the aim of our project is to develop an information platform in which anyone, whether an expert or an ethnic user, can contribute knowledge and skills, misinformation can seriously undermine the reliability and integrity of our system. Many ethnic minority users have low information literacy, making them susceptible to potential misinformation on the platform. Additionally, ethnic minorities tend to live in close-knit communities where misinformation can quickly circulate in person or online. To tackle this issue, this design concept proposes a misinformation detection system that verifies the integrity of the uploaded information before committing it to the database.

2) Description

a) Design outline

To tap into existing ethnic knowledge bases, our information system is designed so that ethnic users can share their knowledge and skills with others. Users can upload their content to the system through a user interface. One major flaw to this level of access is the risk of mis/disinformation being uploaded and consumed on the platform. This design offers a solution to this problem by installing a misinformation detector between the user upload interface and the database, which

ensures that the content is verified before going into the database. This detector deals with video content, which is the predominant kind of content on our platform.

The misinformation detector is based on the framework developed by Shang et al. (2021) [\[25\]](#) that is used to identify misleading COVID-19 videos on TikTok. Their approach is multimodal, meaning that it identifies misleading content based jointly on the visual, audio, and textual information of the videos. The motivation behind this framework is that using any form of information in a video (its visuals, audio, or texts) independently is not enough to identify mis/disinformation because mis/disinformation can be expressed in different interconnecting ways. It is through the consideration of all media as a whole that mis/disinformation can be confirmed.

In this approach, a video is defined as having visual content (which is a sequence of frames), audio content, a description, and metadata. Following their approach, we will run each uploaded video through our misinformation detection module before saving them to the database. The module consists of four major components:

- The first component (Caption-guided Visual Representation Learning) identifies the visual features from the video frames which relate to the topic (which in our case is agriculture) using any text shown on screen and words in the audio around the frame.
- The second component (Acoustic-aware Speech Representation Learning) learns from the audio content to find the set of words present in the speech as well as their corresponding audio segments, which can contain important acoustic features such as volume and tone that indicate key information. The output of this component is a hybrid representation that captures both the words and audio.
- The third component (Visual-speech Co-attentive Information Fusion) fuses the information extracted from the previous two components by learning the relationship between each video frame and speech word, assigning a higher score to pairs that are more closely related in terms of expressing misinformation. The result is a vector representation of each video that can be fed to the next component for classification.
- The fourth component (Supervised Misleading Video Detection) features a neural network classifier that takes in the vector representation of the video and classifies it as either misleading or non-misleading.

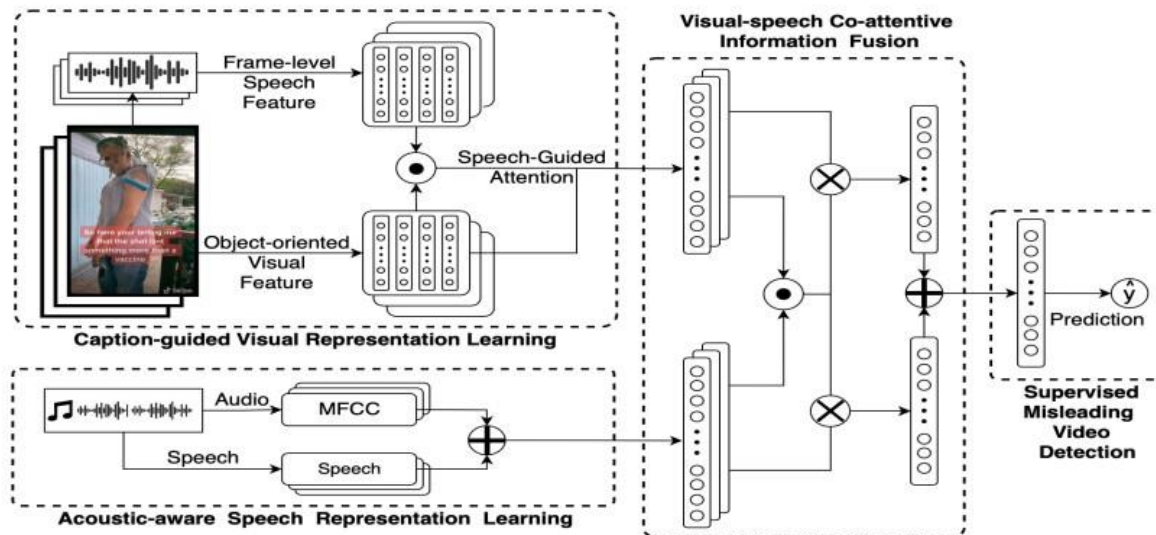


Figure 5. Multimodel for misinformation checking [\[25\]](#)

If an uploaded video is marked as misleading, a notification will be sent to the uploader warning them of the flagged content and the video will not be saved to the database. After enough warnings are sent, the uploader will receive penalties or permanent bans from the platform. However, the model can still occasionally produce incorrect results (i.e. it flags non-misleading videos as misleading). In this case, the uploader can issue a report so that the video can be checked again by professionals for a final verdict. All of this will be achieved through a friendly user interface. The following diagram illustrates how the misinformation detector will be integrated into the structure of the information system:

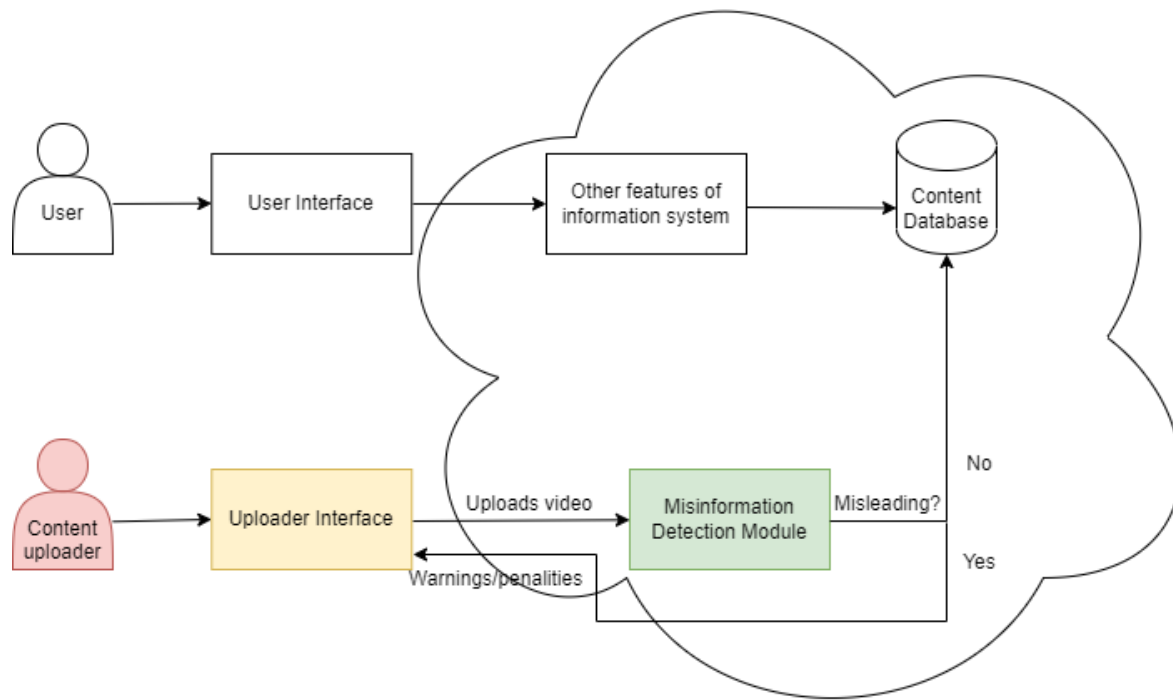


Figure 6. Cloud workflow for misinformation checking.

The structure diagram showing the location of the misinformation detector relative to the other features of the information system.

b) Specifications

To train and deploy the machine learning model for the misinformation detector, this design will leverage the Amazon SageMaker service. SageMaker is a managed service for building, training, and deploying machine learning solutions. We will deploy our solution in the Asia Pacific (Hong Kong) region, which is the cheapest and closest to Vietnam. Since our machine learning task involves natural language processing (analyzing text on the screen and words from the audio) and computer vision (identifying key objects in the video frames), we will leverage the ml.g4dn instance family, which provides the accelerated computing power needed for this workload.

To estimate the pricing for SageMaker, the following aspects will be considered:

- Notebook instance: These are the compute instances where the code to create and train the models will be run. For the ml.g4dn family, pricing starts at \$1.134/hour.
- Training: This refers to the tools used to train, tune, and debug the model. For the ml.g4dn family, pricing starts at \$1.134/hour.
- Inference options: Two inference options are available, which are real-time and asynchronous. Real-time inference enables real-time predictions, which allows users to get

immediate feedback on the validity of their uploads. For the ml.g4dn family, real-time pricing starts at \$1.134/hour. Asynchronous inference queues and processes requests asynchronously, meaning that users do not get immediate feedback. Pricing for asynchronous inference starts at \$2.317/hour but the instance number can be autoscaled to 0 when there are no requests, which might save costs in the long run.

All of the information above is taken from the official website of AWS (<https://aws.amazon.com/sagemaker/pricing>)

c) Benefits

This design concept will ensure that our information system remains a useful and reliable source of information for ethnic minorities looking for agricultural knowledge. It enables the safe exchange of ethnic agricultural knowledge among minorities. Having all of our content verified will prevent users from getting conflicting results when they look up information on the platform. This is especially useful for ethnic users with low information literacy who cannot tell apart right and wrong information.

The misinformation detector will also prevent bad actors from undermining the integrity of our information system by uploading misleading content, which can have seriously harmful effects on communities. The penalty/ban system will prevent repeat offenders from using the system.

d) Constraints

One issue with this design concept is the collection of training data. Before we can deploy the learning model for production use, we have to train it on a dataset containing videos about agricultural topics. We need to collect a wide range of both misleading and non-misleading videos, analyze their information, and pre-label them to prepare for training. As our group members do not have specialized knowledge, we will need to consult agriculture professionals for the collection and preparation of the training data.

Another problem with this design is that the learning model can occasionally misclassify videos. It may label non-misleading videos as misleading and vice versa. This is expected of any machine learning solutions, and rigorous testing will be done to minimize the margin of error. During the implementation of the system, we will provide users and uploaders the option to report content that has been incorrectly flagged so that it can be manually verified by professionals.

V) Design concept 5: Auto translation with Transformer model (Tran Khai Kiet)

1) Learning issue

For a long time, most of the information about agriculture was stored in books and other paper-based media. However, ethnic minorities usually face challenges due to differences in languages between ethnicities. As a result, most of the current vocational training programs or agriculture workshops are out of the scope of those who do not know general Vietnamese.

2) Description

a) Design layout

To cope with this problem, the use of an agriculture knowledge database that is designed especially to support mountainous farmers is required. A core database will be used to store tutorials uploaded to the system but before pushing the data in, the uploaded content will need to be translated during this phase.

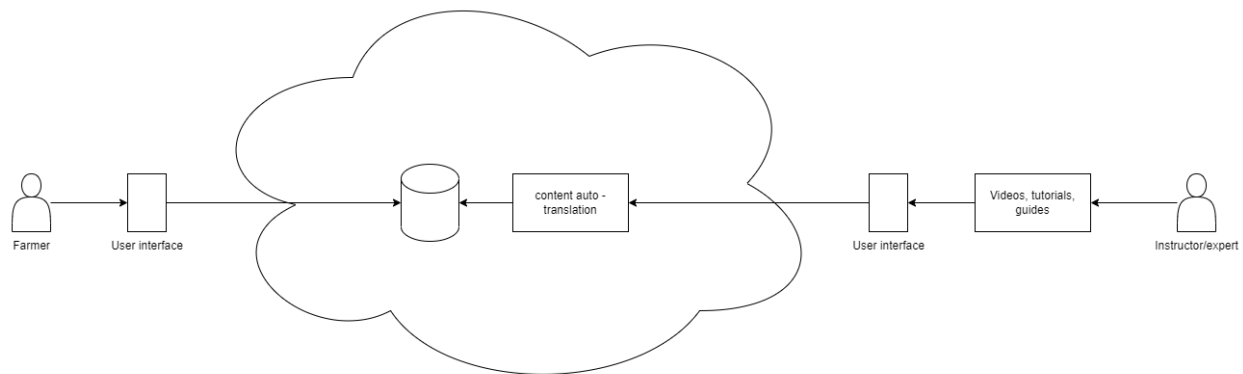


Figure 7. Cloud workflow to translate video content.

For translating the video and tutorial content uploaded to the system, we will use a parallel dual-decoder Transformer to get the transcript of the video while translating it at the same time. The speech of the video will be read in with an input layer and fed into the transformer model. Because this model will simultaneously do 2 tasks, which are speech recognition and translation, the model will need 2 parallel decoders working together to get 2 outputs at the same time and feed it forward. Moreover, it also has an encoder to constantly read words from input data [\[26\]](#).

More specifically, the transformer model will utilize the attention mechanic for NLP. Most of the NLP neural network works on the principle of recurrent neural network to keep the link between input words. However, this model has a critical weakness, which is its performance. The model will run very slowly and inefficiently when being trained on a large dataset. However, the introduction of an attention matrix, which helps the NLP model to keep track of the relationship between words instead of recurrence, has revolutionized the world of NLP. As a result, model performance has become better even when it works on large dataset, thus make room for a better AI model to be trained on large dataset. Unfortunately, even with this technology, the model still only can reach a maximum of 67% accuracy [\[26\]](#). The reason for this is that the model has to do 2 tasks at the same time for video content translation.

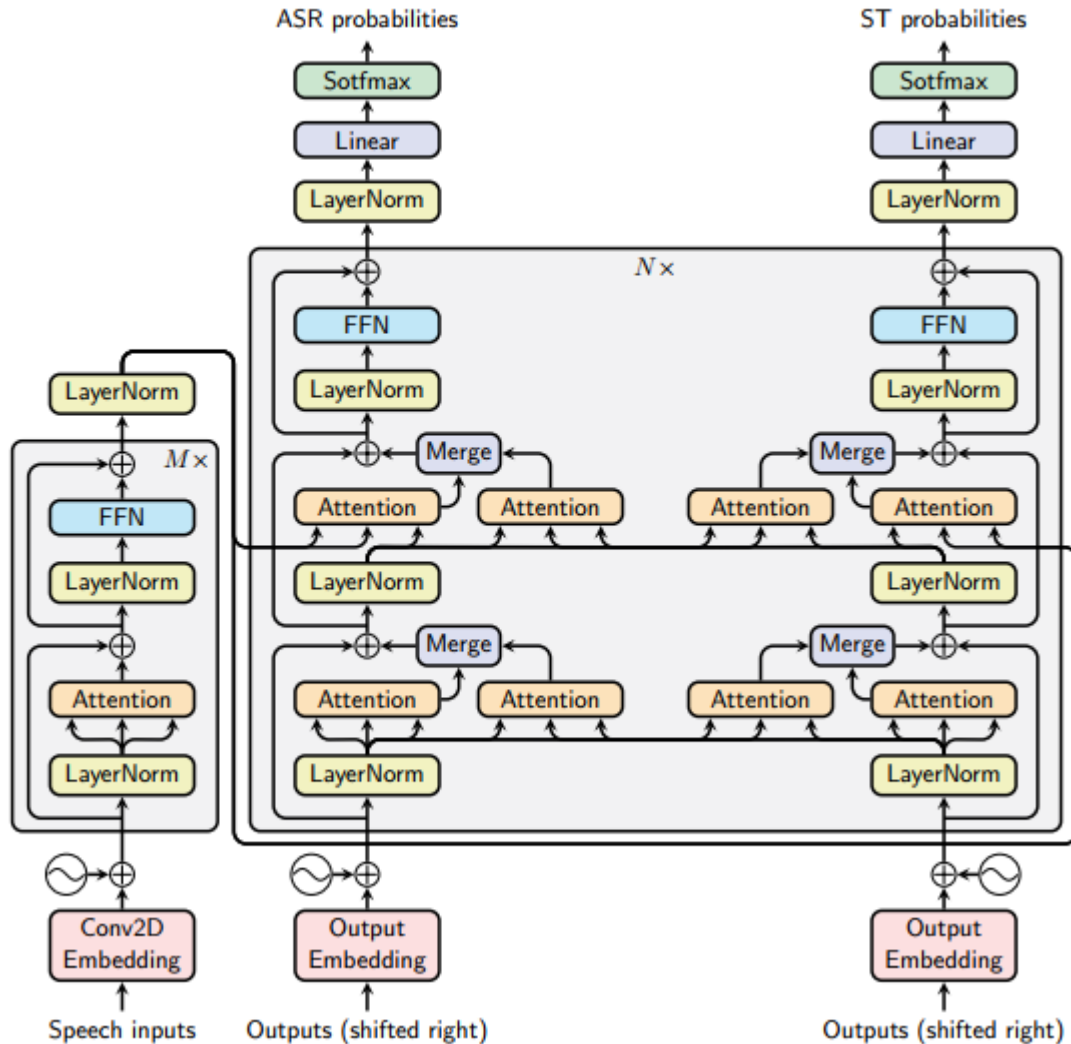


Figure 8. Transformer model for video script translating [26]

b) Specifications

To deploy the desired architecture, a video storage space will be required since it is the core feature of this idea. Moreover, developing an AI model will require a GPU to train the model, and deploying this AI model will also require a proper cloud service.

The table below will list all the required components for this implementation with their specifications.

Components	Name	Cost (let the request used per month = 1000 requests/month)

Storage [27]	S3	$\$0.023 + \$0.005 * 1000 + \$0.0004 * 1000 = \5.423
A database to store references to storage [28]	MongoDB (serverless version)	$\$0.1/1000 = \$0.00001/\text{month}$
GPU	Tesla T4	\$1,187
Deep learning compute unit for deployment [29]	ECS + EC2 (spot) + EC2 (on-demand)	$\$0 + \$25.988 + \$30.368 = \56.356

Table 3. List of components and pricing

All the above components also have many good alternatives that can be taken into consideration. For the storage service, S3 can be replaced with EBS which is another type of cloud storage. However, S3 will be more suitable for this project since S3 is more flexible and cost-effective, which is the top 1 priority for testing and prototyping. Moreover, S3 is also more scalable than EBS since it is not limited by the instance type while EBS does.

For GPU, there are also many other options besides Tesla T4. However, in terms of compute compatibility which is required to train some special Machine learning and Deep learning models, this GPU has a compute capability rate of 75 and will work well on this task. Furthermore, Tesla T4 is also a cost-effective option since it only costs around \$1100 to do the training task at an acceptable rate.

After creating the AI model, deploying it will also require the use of computing units. There are 2 viable solutions for this. The first solution is to use AWS SageMaker which is a pre-built architecture for machine learning and deep learning provided by AWS. However, AWS SageMaker has been reported by some developers as being inefficient. Therefore, for this project, the model will be deployed with the containerization method. This model will be containerized with Docker and then deployed onto AWS with ECS and the support of EC2.

c) Benefits

The development of this database can bring many advantages to ethnic farmers.

Farmers with low literacy will still be able to learn and develop their skills just by accessing this database since it supports multilanguage translation. Previously, most of the available information for mountainous farmers is provided in Vietnamese, which is not always their mother tongue, resulting in the marginalization of these people. However, with this database system, these marginalized people will be considered.

Moreover, the translation system will also help the search engine to help mountainous people to search for specific content in their language. This system will also cooperate with the search engine to check if the queries match not only the content but also the translated version of it, thus allowing the mountainous people to browse the courses and tutorials they need with ease.

d) Constraints

Despite the benefits that this system can bring to ethnic users, there are also many constraints that can impact their experience on the platform.

One of the constraints is the effectiveness of the AI system. As shown in the research paper, the model only has an accuracy of at least 67% at its best. However, to help the ethnic people fully understand the content of the tutorial, the model will need to achieve a better performance than that. Therefore, to further develop this model, the use of other knowledge models to support the main model will need to be looked into.

Another constraint is the complexity of the machine learning preprocessing system. The implementation of these systems carries many potential risks that might cause harm to the overall system since the output of these models will then be used to support other operations of the system. Therefore, the question of compatibility also requires a proper answer before bringing all these ideas into action.

References

- [1] Nguyễn Tùng Phong, Trần Đức Trinh, Lê Thị Hồng Nhung, Lê Văn Chính (2017, August). *Training to apply technology development in agriculture in Viet Nam: issue and solution*. ResearchGate.
https://www.researchgate.net/publication/351449300_Dao_tao_ung_dung_tien_bo_khoa_hoc_cong_nghe_phuc_vu_san_xuat_nong_nghiep_o_Viet_Nam_Thuc_trang_va_giai_p_hap
- [2] Trần Quốc Nhân, I. T. (2012). Analyzing Causes of Failure in Contract Farming Enforcement between Farmer and Entrepreneur in Vietnam. *Tạp chí Khoa học và Phát triển*, 1069-1077.
- [3] Ha Thi Hai Do, C. N. (2020). Impact of Vocational Training on Wages of Ethnic Minority Labors in Vietnam. *The Journal of Asian Finance, Economics and Business*, 7(6), 551-560. doi:<https://doi.org/10.13106/jafeb.2020.vol7.no6.551>
- [4] Vidanapathirana, N. P. (2019). Agricultural information systems and their applications for development of agriculture and rural community, a review study. *Research gate*.
- [5] *Results of the survey to collect information on the socio-economic status of 53 ethnic minorities in 2019*. Ha Noi: General Statistic Office.
- [6] Wenjing Pian, J. C. (2021). The causes, impacts and countermeasures of COVID-19. *Elsevier*.
- [7] Doshi, Z., Nadkarni, S., Agrawal, R., & Shah, N. (2018, August 1). *AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms*. IEEE Xplore. <https://doi.org/10.1109/ICCUBEA.2018.8697349>
- [8] Folasade Olubusola Isinkaye, Yetunde Folajimi, & Bolanle Adefowoke Ojokoh. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3), 261–273. <https://doi.org/10.1016/j.eij.2015.06.005>
- [9] Nguyen, T. H. (2020, May 26). *Đào tạo kỹ năng nghề nông nghiệp: Vấn đề bức thiết*. Tạp Chí Điện Tử Kinh Tế Nông Thôn. <https://kinhtenongthon.vn/dao-tao-ky-nang-nghe-nong-nghiep-van-de-buc-thiet-post35666.html>
- [10] Woodward, R., Romera, A. J., Beskow, W. B., & Lovatt, S. J. (2008). Better simulation modelling to support farming systems innovation: Review and synthesis. *New Zealand Journal of Agricultural Research*, 51(3), 235–252. <https://doi.org/10.1080/00288230809510452>
- [11] Hang Nguyen Thi, Y. N. (2021). Improving the quality of human resources in the north mountainous province of vietnam to meet business demand in the context of digital transformation. *The USV Annals of Economics and Public Administration*, 21(2), 84-92.
- [12] Jun-Ho Huh, S.-K. K. (2020). Verification Plan Using Neural Algorithm Blockchain Smart Contract for Secure P2P Real Estate Transactions. *Electronics*, 1052. doi:<https://doi.org/10.3390/electronics9061052>
- [13] *A Comparison between AWS S3 Infrequent Access and Azure Cool Blob Storage*. (2022, 10 28). Retrieved from Stonefly: <https://stonefly.com/blog/comparison-aws-s3-infrequent-access-azure-cool-blob-storage>

- [14] Rifai, M. (2023, 6 8). *Cloud comparison: AWS EC2 vs Azure Virtual Machines vs Google Compute Engine*. Retrieved from Pluralsight: <https://www.pluralsight.com/resources/blog/cloud/cloud-comparison-aws-ec2-vs-azure-virtual-machines-vs-google-compute-engine>
- [15] *Amazon Elastic Container Service (Amazon ECS) vs. Azure Kubernetes Service (AKS)*. (2021, 9 9). Retrieved from TrustRadius: <https://www.trustradius.com/compare-products/amazon-elastic-container-service-ecs-vs-azure-kubernetes-service-aks#pricing>
- [16] White, N. M. (2022, July 15). The Hmong Medical Corpus: a biomedical corpus for a minority language. *Language Resources and Evaluation*, 56(4), 1315–1332. <https://doi.org/10.1007/s10579-022-09596-2>
- [17] *How to improve search results for your website or app - Every Interaction*. (n.d.). Every Interaction. <https://www.everyinteraction.com/articles/improve-search-results-for-your-website-or-app/>
- [18] (2021, March 17). *Natural Language Processing In 5 Minutes | What Is NLP And How Does It Work? | Simplilearn* [Video]. YouTube. Retrieved June 20, 2023, from <https://www.youtube.com/watch?v=CMrHM8a3hqw>
- [19] M. (2021, September 30). *9 Principles of Good Web Design - read our guidelines to consider*. Feelingpeaky - Creative Design Agency, London. <https://www.feelingpeaky.com/9-principles-of-good-web-design/>
- [20] Shiotsu. (2021, March 1). *What NLP is & a Closer Look at Google Cloud Natural Language API*. Upwork. Retrieved June 20, 2023, from <https://www.upwork.com/resources/natural-language-processing-and-google-cloud-natural-language-api>
- [21] M. (2022, May 4). *A Quick Guide to Low-Resource NLP - MLOps Community*. MLOps Community. <https://mlops.community/a-quick-guide-to-low-resource-nlp/>
- [22] *Comparison of the Most Useful Text Processing APIs | ActiveWizards: data science and engineering lab*. (n.d.). ActiveWizards: Data Science and Engineering Lab. <https://activewizards.com/blog/comparison-of-the-most-useful-text-processing-apis/>
- [23] *Pricing | Cloud Natural Language | Google Cloud*. (n.d.). Google Cloud. <https://cloud.google.com/natural-language/pricing>
- [24] *Amazon Comprehend – Pricing*. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/comprehend/pricing/>
- [25] Lanyu Shang, Ziyi Kou, Yang Zhang, Dong Wang. (2021). A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. *IEEE International Conference on Big Data (Big Data)*.
- [26] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, Laurent Besacier. (2020). Dual-decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation. *arXiv*.
- [27] *Amazon S3 pricing*. (n.d.). Retrieved from AWS: <https://aws.amazon.com/s3/pricing/>
- [28] *MongoDB Pricing*. (n.d.). Retrieved from MongoDB: <https://www.mongodb.com/pricing>

[29] *Amazon EC2 pricing*. (n.d.). Retrieved from AWS: <https://aws.amazon.com/ec2/pricing/>