iVehicle: Vehicle based Natural Language Processing Search Engine

Rathnayake R.M.V.D.
Faculty of computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Ushana W.A.V.
Faculty of computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Fernando W.A.P.N.
Faculty of computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Kumarasiri B.L.B.M.
Faculty of computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

ABSTRACT
People constantly utilize technological advances, such as smart devices, to make their daily lives easier. The majority of those folks prefer to communicate with machines using natural language. The goal of this study is to develop a natural language search engine for purchasing and selling vehicles. When searching using text or voice, the system will provide the best results for the user. There are a few search engines where a user can search for what they need in natural language, yet the user is frustrated since the search engine does not provide the best possible results. A search in natural language, such as English, is conducted using conventional spoken language, such as Google or Bing. Users may be necessary to visit a variety of different websites and search engines while searching for vehicles and parts. When browsing for vehicles, the searcher is usually presented with a list of records rather than genuine search results due to inaccuracy and duplication of search results. People's preferred search engine was frequently interfering with their own internet results, causing annoyance, and recurring difficulties with online search vehicles. The reason for the delay in delivering results is partly due to the volume of results that a natural language search is expected to return. This approach tries to make it easier for users to access information about vehicle purchasing and selling, as well as facts about the vehicle price process, through a specially designed mobile and web application. This system, which will be developed in both Sinhala and English, has been designed. Because of the illiteracy of the population, that system was created to conduct voice searches. Machine learning technology will be used to create this system.

Keywords
Machine learning, search engine, natural language processing

1. INTRODUCTION
Consumers in obtaining relevant results depending on their search field, Internet search engines with access to a significant amount of data were developed. At the moment, search engines are one of the most popular web applications [1]. They function as a scanning system for the contents of other systems in order to locate information on the internet. Among the most popular search engines are Amazon, Bing, Google, and Baidu [2]. They provide superior search capabilities by indexing a large number of web pages.

Certain search engines specialize in particular fields, such as Google Scholar, which is devoted entirely to research, publications, and articles. Students and researchers can use Google Scholar to conduct research relating to their study.

Similarly, this research will concentrate on vehicle and vehicle components. Natural Language Processing (NLP) is the cornerstone for this project, which will provide information about autos and vehicle parts to humans who express an interest in them. The vehicle is a sophisticated electronic system with numerous components with a variety of names that can be spoken and written in a variety of languages [3]. When individuals seek for vehicles or vehicle parts, they are mostly concerned with the brand name, color, manufacture year, and a few other characteristics.

At the moment, there are a few solutions available for purchasing or selling vehicles based on their manufacturing characteristics. They have already incorporated some of the techniques discussed under Natural Language Processing into these systems. Humans, in particular, employ a variety of systems to meet their needs and desires. Generally, customers prefer to obtain the optimum outcome from a single integrated solution. When users enter all of the vehicle's or component's details into the system, existing systems return a plethora of results in addition to the user's requirements. As a result of this research, a search engine system based on Natural Language Processing will be developed by combining the majority of web platforms and mobile applications that are utilized by users to search for vehicles that meet their specific requirements. Though users must submit their own impressions when searching for details about vehicles or vehicle parts, the systems will be unable to meet their expectations every time. As a result, by focusing on NLP, this research hopes to provide the user with an accurate result that meets their needs [4].

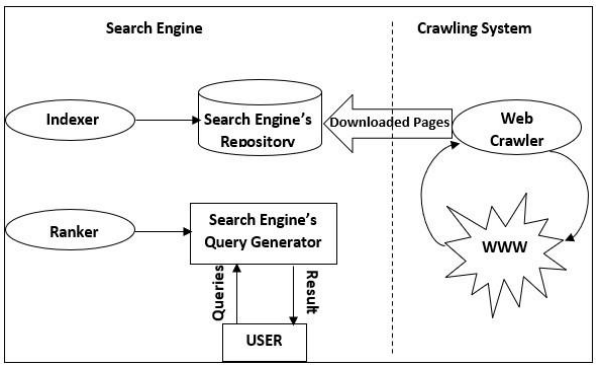

Fig 1: Process Overview Diagram of the existing system[5]

Existing search engines have amassed vast collections of web pages that have been crawled. Once users begin looking for or gathering information relevant to their jobs, the system is

utilized to save the links to the web pages that the user is typically searching. As a result, if the user attempts to obtain information on a particular area, the system will be able to filter the details from the links that the system has previously recorded.

The crawler attempts to locate the desired information by extracting the links in this case. In this situation, the crawler concentrates on the fewest possible keywords and tries to provide all the details associated with those keywords. As a result, there may be irrelevant data that is unrelated to the requested field.

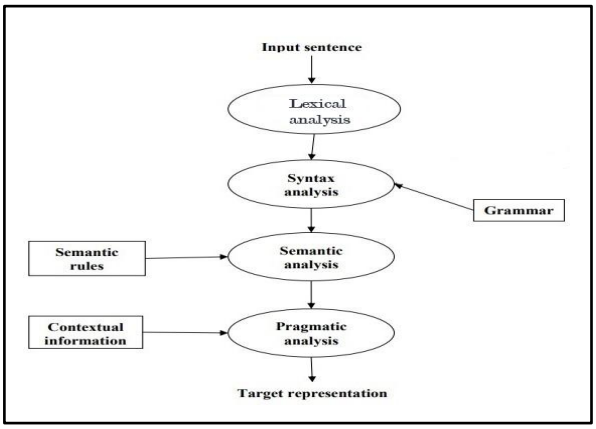

Fig 2: Flow Diagram of the proposed system

As a more efficient and effective replacement for the current method, the suggested system utilizes Natural Language Processing to enhance the process. The suggested system is designed to accept input sentences and follow the flows depicted in Figure 2. After obtaining the input sentence and correcting the grammar, the suggested system will concentrate on identifying more key terms from it. The crawler will then be able to filter out the material that is most relevant to the user. As a result, the proposed system's accuracy can be increased. Additionally, the suggested system would consider the Sinhala language through Natural Language Processing. This will be more user-friendly for Sri Lankans in particular. Present, there are no automobile specific search engines in Sri Lanka. Natural Language Processing and voice search are not supported by the majority of search engines. As is customary, customers must first navigate through the individual systems and then to the top advertisement or new advertisement while searching for vehicles. They are not going to offer their own concept a favorable first impression. As a result, the research paper will consider these issues.

In this research related works will be discussed in the literature survey and in the next section tools and technologies are discussed under methodology. Then, it will ensure the results of the proposed system and finally it will focus on the future developments.

2. LITERATURE REVIEW

The Natural Language Processing concept has been growing globally. Natural Language processing can be used to extract important data from the user inputs. Nowadays people mostly used to accomplish their needs and wants over the internet. Because of these kinds of aspects most people are used to selling or buying vehicles and vehicles related parts over the internet. Most of the people are willing to buy/sell vehicles over the internet. Though there are natural language search engines, the expected results are not accurate, and a bunch of unrelated data have been triggered.

At the moment, various types of search engines exist for a variety of distinct jobs. At the moment, the most popular search engine is Google, but there are others such as Bing, Yahoo, Baidu, and AOL [6]. However, there is no such search engine dedicated to the automobile industry. The research conducted the investigation using a search engine tailored to the vehicle industry. Additionally, this search engine supports the Sinhala language. Rather than that, users can search using both text and voice inputs. Natural language and voice input are not supported by the majority of search engines. As shown in the below table it illustrates our proposed system and the other existing systems and their features.

Table 1. Table of comparison of the proposed system and previous research

Features	Analysis of the Combination of Natural Language Processing and Search Engine Technology [7]	Natural Language Processing in Information Retrieval [8]	Our System(“iVehicle”)
Chatbot assistant.	X	X	✓
Voice search for support for Sinhala and English languages.	X	X	✓
Search based on Vehicles.	X	X	✓
New result notification.	X	X	✓
Gather user information.	✓	X	✓

In this system there is a chat assistant as well. According to the technologies used, chat assistants are divided into two main categories. There are two types of chatbots: rule-based and AI chatbots. Prior to the era of AI, chatbots were created using rule-based approaches. With the evolution of AI and natural language processing, the majority of modern chatbots that are utilized for human-to-human contact are built using machine learning technologies. Numerous systems are available for implementing conversational agents. These technologies are offered by a variety of solution providers and provide a variety of characteristics.

Mainly machine learning based chatbots use CNN, LSTM, ANN and RNN. Google DialogFlow [9], IBM’s Watson Assistant, LUIS.ai [10], and Wit.ai [11] are the primary pay

as-you-go chatbot development frameworks, whilst the RASA framework is one of the easiest and famous opens source framework that helps businesses improve their conversation and interaction with their customers or audience [12].

Table 2 compares existing chatbots in terms of capabilities, supported languages, and simplicity of use. Developers who seek to construct chatbot models using native languages or multilingual training examples will need to employ a framework such as RASA in order to maintain maximum control over their chatbots. Mainly Rasa Supports Sinhala language as well. Below table depicts the supported languages and features.

Table 2. Chatbot Frameworks

	Ease of Use	**Features**	**Languages**
Dialogflow	Provides a web interface	Basic inbuilt web integration	Supports 20+ languages including English, Spanish.
	For creating robots which makes it easy for anyone to create basic bots.		Portuguese, French, Hindi, Chinese etc.
Amazon Lex	Provides a web interface for creating and launching robots, run on the same machine learning engine as Alexa	Basic chat UI provided for testing on the website.	Currently, only US English is supported.
IBM Watson	Provides a good and easy to navigate user interface, Readymade samples and videos available for quick start.	Basic chat UI for websites.	Support 10+ languages (in BETA) including English, Spanish, Japanese, Italian, Chinese etc.

Table 3 compares the fundamental functionality provided by the various services. All of them, with the exception of Amazon Lex, are based on the same fundamental concept: The user can train a classifier on example data to categorize so-called intents (which indicate the overall intent of the message and are not confined to a specific point within the message) and entities (which can consist of a single or multiple characters).Below table shows the comparison of basic functionality of NLU services. [13]

Service	Intents	Entities	Batch import
LUIS	+	+	+
Watson	+	+	+
API.ai	+	+	+
wit.ai	+	+	0
Lex	+	0	-
RASA	+	+	+

Fig 2: Flow Diagram of the proposed system

Chatbots are applicable to a wide variety of different domains. Apart from chatbots for personal use such as Google Assistant, Siri, and Amazon Alexa, chatbots are also utilized in corporate domains such as banking, travel, fashion, and recruitment. For instance, Sampath Bank [14] and American Express's Amex bot [15] are two robust chatbots accessible in the banking sector. Sampath Bank chatbot is one of the first artificial intelligence-powered banking chatbot made in Sri Lanka, providing quick responses to millions of consumer enquiries across numerous platforms. At any moment, the Amex bot delivers information about a customer's purchase history made with an American Express card. Expensify: "Concierge" [16] is a chatbot used in the travel sector to assist businesses across industries with streamlining and automating expenditure reports and trip plans. The Official H&M Chatbot [17] is said to be one of the greatest in the fashion industry, providing customers with a fully functional online fashion retail experience. Mya [18] is a successful artificial intelligence (AI) recruiting assistant at First Job. It maintains enormous candidate pools, freeing up recruiters and hiring managers to focus on interviews and closing offers. The iVehicle chatbot provides information about vehicles and also supports the Sinhala language.

Chatbots are gaining popularity rapidly due to their performance, speed of response, and, most importantly, their high availability. Nowadays, more businesses are incorporating chatbots into their operations and making their services available 24 hours a day. Additionally, when compared to human power, chatbots significantly boost a business's reach to clients.

3. METHODOLOGY

The proposed “iVehicle” is a search engine system that is specialized in vehicle searches. This should be capable of, Real-time capturing the user’s voice and convert it to a text base output, implement a chat bot to assist the user and develop a mobile application and web application. Natural Language Processing (NLP) plays a big role in day-to-day life. An Artificial Intelligent function processes user input natural language data and gives the output as the user needs. There are four main functions in natural language processing.

To enable the search engine “iVehicle”, the system is designed to occupy a mobile application developed for the usage of the vehicle ad searcher to buy or sell vehicles. The system overview diagram for the proposed system is depicted through figure 3.

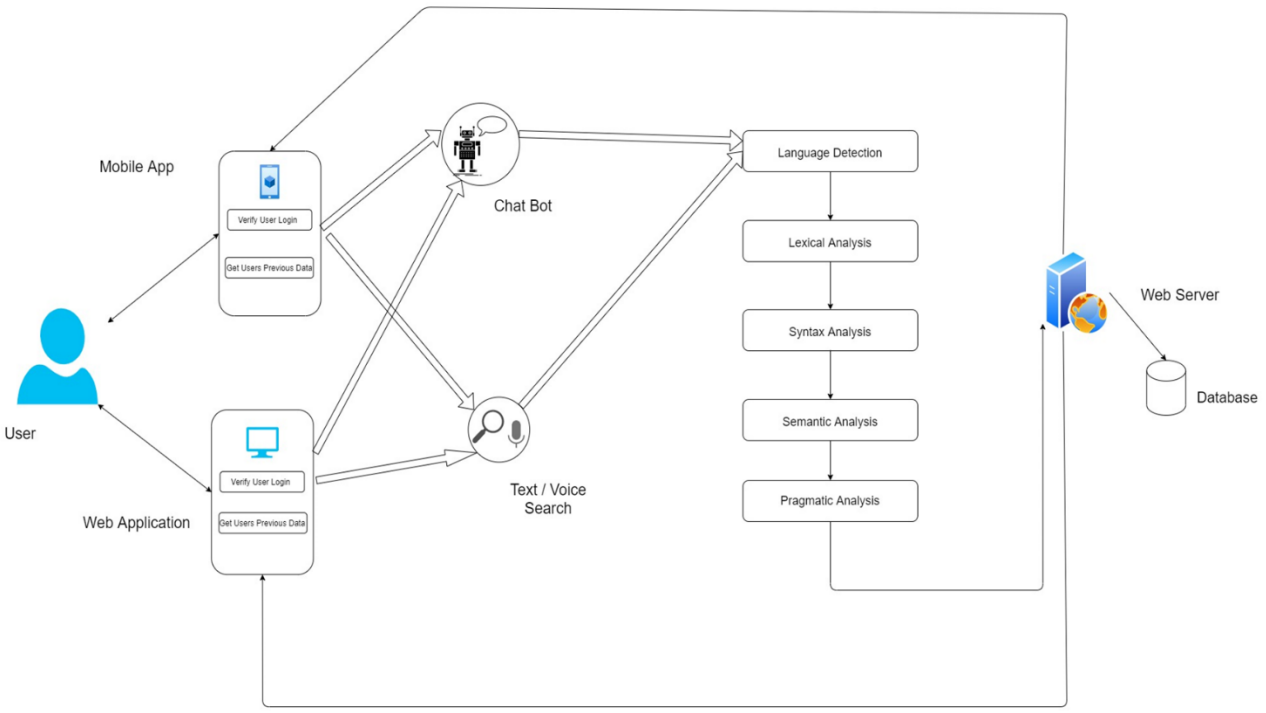

Fig 3: System overview diagram of the proposed system

3.1. Accumulating a user's search

Through the mobile application, once the user is logged successfully into the system, the user will be able to navigate through the selection of options either to select search predictions as the user's previous searches or else start a new search to search about vehicle ads to buy or sell. In the segmentation of search predictions as the user's previous searches, the system will be creating a user's profile when the user signs up for the system. Then the system will keep data such as what ad type the user searches mostly, what is the vehicle that user searches mostly. After analyzing those data, the system will provide some of the searches in the beginning before the user starts a new search. Once the user selects search predictions which is popup to the user, the prediction number will go to the system and check the previously searched item number and get the results through the backend and give the same results in the results view page as listed card views. From that the user can select the search result and view the full ad on the result.

When the user tends to proceed with a new search, the user can simply type the need in the search bar and press the search button to view the result. If any user is literally disable in how to use mobile apps and type, the user can simply press the voice command button and tell the system what is the need. If it is a text search, the system will process the data using machine learning and get the idea of what is the need of the user and predict results. If the input is voice to the system, first the system will convert the voice to text and then do the same thing that the system does to understand the natural language using machine learning and output better results for the user. If the system looks so complex for some users, they can use the Artificial Intelligent driven chatbot assistant and get help on how to use this system. Other than that, the user can easily use the chatbot to find results.

3.2. Process the users input

Natural Language Processing relies heavily on lexical analysis. Most research did lexical analysis on user input text, but there was no voice and lexical text analyzes were equally unprecise. System must collect text or voice from the user input, transform it to texts and construct suitable words and phrases with the help of machine learning. In the initial phase, collected data for lexical analysis from an existing dataset. Next to achieve a better outcome, have preprocessed the data. Then the categories were used to organize the data into distinct words and phrases. Next, as data items, the outcome had many groupings. The sort of the data types was examined to learn more.

Next, the label encoder was used to convert strings to integers. converting the strings to integers is needed because those integers will help to get a better result accuracy when using machine learning algorithms. Finally, supervised machine learning algorithms were used to check, which one gives the better result accuracy. Classification and Regression Trees were used as the machine learning algorithm. Other than the text input, voice input was considered for the lexical analysis. To do that an IBM API was used to convert voice data to text in the English language. Sinhala Language was also considered as a new thing when doing this research. Google has provided an API called Google Trans to convert the Sinhala language to the English language. If the user input was in Sinhala language, the Google Trans API will convert it into English language and pass that output to do the lexical analysis. If the user's input is in Sinhala Language voice, then the "CMUSphinx" library was used to convert the voice to text and then the converted text was sent as the input to the Google API to translate to English. Then the same process which has been done to the text input will be done to the voice input to do the lexical analysis.

The next important element of the process after lexical analysis is syntax analysis. The main aim of Syntax Analysis is to rectify the grammar that the lexical analysis yielded. To do that, an existing data set with grammar corrected sentences based on vehicle search was used. Next, use a label encoder to convert the strings to integers. Then, some unique data was found from the dataset. After that, Machine Learning algorithms were used to get a proper result when the input is

grammatically correct. In the next step, a grammar rule was created to find the grammar incorrect sentences and correct them. To do that, tree structure was used and got an idea how a sentence places the word types.

Structure of the sentence parsing is often depicted in Figure 4 as a root-based parsing tree. Three nodes are present in the parse-tree, which are terminal nodes at the bottom of a tree with the word, pre-finish nodes directly connecting them to each terminal node carrying a phrase or a higher syntactic structure, and non-terminal nodes.

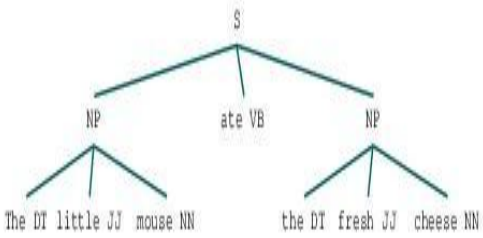

Fig 4: Tree Structure of a sentence

The system can detect nouns and verb-phrases to identify probable subjects, predicates and objects in the sentence. This information is required to infer meaning from phrases and may be utilized to comprehend a text-corpora context.

Semantic analysis is a major element of the processing of natural language. It offers us the wording of the dictionary from the grammar rectified by syntax analysis. For the actual meaning of the sentences or words, the system should capture the result of speech analysis, convert it to text and collect the information. In present systems, while looking for a car, consumers don't obtain precise, useful dictionary results. Use long-sentence then semantical analysis to discover keywords in the text, so users require a detailed look at anything. In one word, there are a lot of synonyms. Semantic analysis then finds the synonyms based on the corresponding terms in the vehicle.

First stage, an existing data set was collected to analyze the semantic in order to stop the dataset with the number of words. Then the data set is preprocessed for better results and word bags appear in the dataset comprising the number of time words and extremely unusual and frequent words are deleted. Then, another existing dataset was gathered with keywords based on vehicle searches. The terms in this subject and their proportional weight in each subject. Identify the word synonyms in the dataset. Finally, a machine learning algorithm was trained and obtained the most correct data utilizing the decision tree.

An algorithm that measures how often the term appears in a document is a "bag of words" model. The model's aim is to represent a text as a multiset, that means, that the word order is ignored and the grammar is discarded, while its multiplicity is preserved. The word counts allow for the comparison and calculation of documents, but no semantic structure is saved. The concept is used for applications such as search engines or documents.

Term Frequency-Inverse Document Frequency is an inverted model bag-based weighting technique. Term Frequency-Inverse Document Frequency is employed in the information recovery which, based on the term frequency and inverse document frequency, gives weight for each terminology in a text. If a word is measured in a certain text and common terms are often found in all papers, they should be less valuable than uncommon words. This may be resolved by computing the inverse document frequency, meaning the more papers a term occurs, the less value. After doing Semantic Analysis the next major step in Natural Language Processing is Pragmatic Analysis. Which is giving the output as the user needs it. The system will check the extracted keywords from the semantic analysis and use it as a search key. Then the pragmatic analysis will output the results to the user.

Finally, the chatbot assistant will help the user to use the system. The first part is for training the model, and the second is for testing the model that has been trained. To use the Mozilla Deep Speech model a virtual environment is needed. Then the necessary dependencies can be installed for training the model. Following the steps outlined above, this model should be capable of capturing the user's voice and producing text-based output.

It is necessary to first define the concepts of chatbots and the technologies that will be used to implement one. The two kinds of chatbots are rule-based and self-learning. In the first approach, the bot is trained using rules. A simple question may lead to a simple response from a bot. However, this topic cannot respond to complex queries. For this reason, the research concentrates on self-learning models, such as the Retrieval-Based Model, in which the best response to a question is selected from a list of possible responses or in which the responses are generated autonomously by the bot on its own. The type of natural language should be identified to generate an effective conversational engine, then identify the user's underlying intention, and understand the target audience, how the response should be, and lastly, this research should utilize the language which can answer their questions. Having considered the above information, following are some of the possible technologies that could be utilized to implement the chatbot.

3.2.1. Chatterbot

Chatterbot is a Python-based machine learning-based conversational dialog engine that can elicit responses based on collections of previous conversations. It requires an artificial Chatterbot instance that can be trained and improved using a collection of relevant data that are the objectives, to assist the citizens depending on the issues they face.

3.2.2. NLTK (Natural Language Toolkit)

Natural Language Toolkit (NLTK) is a collection of programs and Python libraries that provides a unique way of working with language-specific statistics. It is one of the most influential NLP libraries which is considered to be the mother of all NLP libraries, with packages for teaching machines to learn human language and respond appropriately. With a series of tokenization, tagging, semantic reasoning, and parsing libraries, like NLTK, the NLTK provides an easy-to-use interface to corpora and lexical resources like WordNet.

It may be possible to employ one of the following suggestions: from the above list, the most appropriate NLP solution will be implemented. To test and train the model, a set of known phrases that may deliver as possible inputs shall be established. Once the model has been developed, it shall be hosted and will be interfaced through the conversational component of the proposed mobile application.

3.3. Giving a user-friendly output

As discussed in the last sub chapter, Natural Language Processing steps will identify the keywords and those

keywords will help to give the results for the user. Google has introduced a programmable search engine with JSON. If we are developing a search engine, it's a waste of time to build a new search engine like Google and predict results. To prevent that kind of issue, NLP controllers as the keyword generator can be used and the google API to search results using those keywords. Next, the results will be called from the system to the application results screen, and it will be viewed by the user. Next, the user can scroll down the results screen card views and select any result to get the full view of it. As the previous chapter mentioned, this is a platform where a user can search for vehicle ads to buy or sell. This platform has included all the vehicle ads-based websites and the results will be shown to the user according to the user's usage history of the system. Users can create a favorite website such as rating, then the results will be ordered to that rating list. Next, there is a notification system to notify the user about a new result for the user's previous search. The notification system can be either ON or OFF as the user needs. Finally, the user can give the rating about the system, if the search is accurate or not.

4. RESULTS AND DISCUSSION

This chapter focuses on the work, analyzes it, and considers the weaknesses of the existing system. Users are not satisfied with the existing system, according to data collected and analyzed throughout the survey. Most of them are willing to adopt a new system based on modern technology. As a result, iVehicle developed a system that allows all applications to communicate with one another. This system aids in the speeding up of the process and primarily aims to be a vehicle-based search engine, with all vehicle-based websites connecting to obtain car information from a single search engine. Furthermore, the user must log in to the application in order to get a search result.

Following the steps shown in the last chapter, classification results were obtained for lexical analysis. The accuracy reached its top value of 0.97 using CART. Linear Regression also reached the accuracy of 0.45. but SVM, KNN and NB got lower-level accuracy in the resulting output 25%, 35% and 37% respectively. The below figure represents the result outcome percentage using machine learning algorithms for lexical analysis.

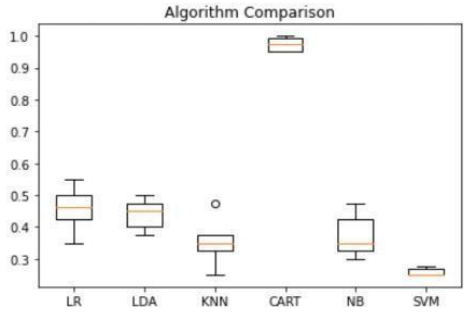

Fig 5: Algorithm Comparison for lexical analysis

In this research the categorization results are taken by the processes outlined in the previous chapter for semantic analysis. Using CART, it shows the accuracy reached a maximum of 75%. Linear Regression also reached an accuracy of 52%. but SVC and KNN got lower-level accuracy in the result output 2% and 73% respectively.

```
LogisticRegression 0.52
KNeighborsClassifier 0.73
SVC 0.02
DecisionTreeClassifier 0.75
RandomForestClassifier 0.75
```

Fig 6: Algorithm Comparison for semantic analysis

To implement the chatbot stochastic gradient descent algorithm is used and after training the model it got 98% accuracy. Also, in my other sub function the decision tree classifier is used to train the model and the bagging and boosting methods are used to increase the accuracy of the model.

The following table represents the queries and the results using the NLP engine.

Table 3. Testing and results outcome

Query	NLP Outcome
I want to buy a black color Toyota corolla which made in 2008	buy, black, toyota, corolla, 2008
I'm looking for a blue honda civic which made in 2018	Looking=buy, blue, honda, civic, 2018

5. FUTURE RESEARCH AND CONCLUSION

This chapter analyzes the work and provides recommendations for system upgrades and future enhancements. There are no other official systems because this system is based on a new concept. As a result, there could be some practical issues when using it. Because this is a fully software-based and smart device-based system, building a user-friendly and extremely simple interface to make the system easier to understand was mandatory. Another concern was the technology/language utilized for the web and mobile apps, which use Angular and React Native, respectively. Attempted to include as many sentences as possible in the research, however there are many types of sentences that were not included. As a result, improve on this work by developing an algorithm that generates a new rule for any new type of sentence that isn't covered by the existing rules.

The proposed approach is designed to assist customers who are unable to locate a vehicle using the English language. Those who do not speak English well might look for vehicles using their native language. Additionally, this approach aims to improve the voice-based search system. This system prioritizes all vehicle applications in order to obtain vehicle information, application usage speed, user-friendliness, and reliability, as well as to assist users. with less technical knowledge in using voice search and chatbots. In the future, the system could be totally digitized, allowing consumers to benefit from new technologies.

6. REFERENCES

[1] S. Bal, "A Comparative Study of Traditional Search Engines with the Metasearch Engines," 2009.

[2] R. Patel. [Online]. Available: https://www.spaceo.ca/ai-

chatbot-development-using-rasa-reasons/.. [Accessed 28 08 2021].

[3] A. A. a. F. H. H. Wang, "Improving Efficiency of Customer Requirements Classification on Autonomous Vehicle by Natural Language Processing," 2020.

[4] M. J. Cafarella, "Research Gate," April 2006. [Online]. Available: https://www.researchgate.net/publication/261860756_BE_a_search_engine_for_NLP_research.

[5] S. Sharma, "Crawler, A Novel Architecture of a Parallel Web," 2011.

[6] D. Dwyer, "Inspire Digital Agency," 1 11 2016. [Online]. Available: https://www.inspire.scot/blog/2016/11/11/top-12-best-search-engines-in-the-world238. [Accessed 28 08 2021].

[7] X. Yuea, "Analysis of the Combination of Natural Language Processing," Elsevier Ltd, 2012.

[8] S. Feldman, "NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval," 1999.

[9] [Online]. Available: https://dialogflow.com/. [Accessed 28 08 2021].

[10] M.Luis, "Microsoft Azure," "LUIS (Language Understanding) – Cognitive Services – Microsoft Azure", [Online]. Available: https://www.luis.ai/home.

[11] "Wit.ai," [Online]. Available: https://wit.ai/.

[12] R. Patel, "Space Technologies," [Online]. Available: https://www.spaceo.ca/ai-chatbot-development-using-rasa-reasons/.. [Accessed 29 08 2021].

[13] D. B. A. H. Mendez, "Evaluating Natural Language Understanding Services," Munich.

[14] "Sampath Bank," [Online]. Available: https://www.sampath.lk/en/personal/electronic-banking/banking-robot.

[15] "Frequently Asked Questions About the Amex Bot for Facebook Messenger," [Online]. Available: https://www.americanexpress.com/us/facebook-messenger/faqs.htm.

[16] D. Faggella, "7 Chatbot Use Cases That Actually Work," [Online]. Available: https://emerj.com/ai-sector-overviews/7-chatbot-use-cases-that-actually-work/.

[17] ""H&M: official chatbot review", 2016,," [Online]. Available: https://bot-hub.com/reviews/official-chatbot-review.

[18] "Mya Chatbot Intro," [Online]. Available: https://www.mya.com/.