Chapter 9
Predicting the Future Research Gaps Using Hybrid Approach: Machine Learning and Ontology – A Case Study on Biodiversity

Premisha Premananthan
Sabaragamuwa Univeristy of Sri Lanka, Sri Lanka

Banujan Kuhaneswaran
https://orcid.org/0000-0002-0265-2198
Sabaragamuwa University of Sri Lanka, Sri Lanka

Banage T. G. S. Kumara
https://orcid.org/0000-0003-3941-2275
Sabaragamuwa Univeristy of Sri Lanka, Sri Lanka

Enoka P. Kudavidanage
Sabaragamuwa Univeristy of Sri Lanka, Sri Lanka

ABSTRACT

Sri Lanka is one of the global biodiversity hotspots that contain a large variety of fauna and flora. But nowadays Sri Lankan wildlife has faced many issues because of poor management and policies to protect wildlife. The lack of technical and research support leads many researchers to retreat to select wildlife as their domain of study. This study demonstrates a novel approach to data mining to find hidden keywords and automated labeling for past research work in this area. Then use those results to predict the trending topics of researches in the field of biodiversity. To model topics and extract the main keywords, the authors used the latent dirichlet allocation (LDA) algorithms. Using the topic modeling performance, an ontology model was also developed to describe the relationships between each keyword. They classified the research papers using the artificial neural network (ANN) using ontology instances to predict the future gaps for wildlife research papers. The automatic classification and labeling will lead many researchers to find their desired research papers accurately and quickly.

DOI: 10.4018/978-1-7998-7258-0.ch009

INTRODUCTION

Wildlife Protection is a trending topic all over the world till the date. Wildlife is critical for the sustenance of life on earth. Conserving biodiversity is critical to maintaining a healthy ecological balance in the world. Especially Sri Lanka has a larger set of biological hotspots which contain a rare and wide variety of fauna and flora. However, the Sri Lankan wildlife is critically threatened due to many reasons, mainly human interventions, and needs dire conservation measures. Lack of Wildlife conservation practices is also impeded by information and technological assistance. Study results of wildlife studies can, but the contribution currently made cannot be satisfactorily integrated into data-oriented conservation and management decisions. This research demonstrates a new method for data mining to find secret keywords and automated labeling for past research activities in this area. Despite the importance and opportunities for the best data resources, there was a lack of research interest from outside of the field. So the problem arises with the retreat to finding research gaps and ideas. Our study focused to sort out this issue while we proposed a method using Machine learning and Ontology to find research gaps and an automatic classification system using past research papers on the Wildlife of Sri Lanka.

The selection of research topics is often not compatible with the actual research needs due to multiple reasons. This is a disheartening scenario as there are plenty of opportunities for such work. Inadequate knowledge of the existing research and its applicability, inadequate use of technology, and inability to locate some research are some of the contributing factors. Other than the research published in a known journal, some past research information available online cannot be found properly because they belong to conventional archives, unfortunately. Increasing public awareness of the values of wildlife and the consequences of losing this heritage can assist conservation to a large extent. To achieve this, we have to simplify the gap between the public and the accessibility to information on wildlife. Technology can play a major role in filling the gap between them. But interacting between the domain and technical party is the problem in our case.

Despite its small size, Sri Lanka has a wide ecosystem diversity because of its topographic and climatic heterogeneity as well as its coastal consequences. So as Sri Lankans we must aware of the wildlife around us. Nowadays there are a lot of crimes and careless attention towards the wildlife especially the rare species. The major reason is the lack of awareness and knowledge. But there is a big number of researches conducted on wildlife. But those researches (Mateo, Arroyo, & Garcia, 2016) didn't reach properly to the outsiders. Mostly the professionals who were involved in wildlife only knew about those. Even other department graduates didn't acknowledge the endangered species or the current situation of our wildlife. This is a bad sign for our country which has one of the world's largest wildlife populations. A small number of online availability of researches were found too. Either there's no proper way to find the availability of past, ongoing research details manually also. So we have to deviate past research techniques to come up with our final solution. Some trending techniques are used here to improve the outputs. Our research aims to resolve the inadequate application of wildlife research and technologies in the decision-making process.

These problems lead to struggle researchers to select this field even they have interests in this field. Mostly wildlife studies aimed to understand species diversity, behavior, and habitat use, and ecology, the role of wildlife in disease transmission, species conservation, population management, and methods to control threats to diversity. In our study, we concentrate on reviewing past research papers using data mining techniques to provide potential research ideas that can be conducted in the future. To fill the data needs for conservation our solution focuses primarily on semi-automating the finding of research gaps

through abstract analysis. Finally, the model includes the most commonly used keywords and question top. This will be a vital milestone for researchers as well as wildlife activists to give an eye on recent problems that need a solution urgently. We must use our data stores efficiently to remove the barriers to easily find research ideas and desired gaps. In this motive, we proposed a novel methodology using trending technologies to make our model more efficient.

There were 3 interrelated steps. First, we collected the past research papers from online and offline in the wildlife of the Sri Lankan domain as much as possible. For the online, we used online resources and offline data from the Department of Wildlife Conservation of Sri Lanka. Here the Latent Dirichlet Allocation (LDA) Algorithm (Movshovitz-Attias & Cohen, 2015) a Bayesian-based Topic Modelling Techniques used to extract the major keywords of each paper from the abstract after that those keywords cluster into each topic with the help of similarity between them. Using those keywords, we modeled the ontology for the prior research works domain. This extracted the specified classes of this wildlife domain of Sri Lanka. Using this Ontology and LDA outputs we performed an intermediate evaluation for labeled the data set and found hidden keywords, creative topics, and intertropical distance also. Finally, the labeled dataset was used to perform Neural Network Classification using Long Short-Term Memory (LSTM) network to train and validate the model. Our study has shown a model for wildlife in Sri Lanka's domain to automatically label, Classify and extract hidden keywords from past research papers. This model accuracy is 83% with the limited dataset. It's a pretty good number if we add more data, we will improve the accuracy at max. The purpose of this chapter is to provide you with a platform that helps to know about hidden keywords and automatic labeling of wildlife researches in Sri Lanka.

BACKGROUND AND LITERATURE REVIEW

From a technical point of view, previous works (Zhu, Klabjan, & Bless, 2017) have shown hierarchical relationship-based Latent Dirichlet allocation (hrLDA), hierarchical topics data-driven model to retrieve terminology ontology from a large number of amalgamated papers, has concentrated on many previous works to come up with a novel idea to solve past problems about automatic classification and labeling. In comparison to conventional topic models, instead of unigrams, hrLDA relies on noun phrases, deals with pattern and text structures, and enhances topic hierarchies with topic relations. Via a series of analyses, we found the Excellency of hrLDA over current topic models, especially for the construction of hierarchy.

Another paper (Singh, Markou, & Haddon, 2000) has shown an efficient approach for training a Neural Network (NN) model to measure moving objects in a video. For the network to simultaneously prepare a named dataset for the first one, object recognition, tracking, and counting are required item identification, pursuing, and counting is threats for efficient Intelligence Transportation Systems (ITS) that used to decrease congestion and detect traffic offenders on highways and in cities. The labeled data creation to training a NN is one of the essential prerequisites for the successful implementation of supervised machine learning.

Medical text classification is considered a special variety of text classification. Health history and medical literature are kept in health documentation, these are the vital instruments for clinical information. In this paper, a unified NN method (Qing, Linhong, & Xuehai, 2019) is proposed. In the sentence representation, the convolutional layer extracts characteristics from the phrase, and both the previous and subsequent sentence characteristics are accessed by a Bidirectional Gated Recurrent Unit (BIGRU). A phrase representation scheme with the essential word weights is used. The process used BIGRU to

encode the sentences received in the phrase representation and decode them so that meaningful phrase weights are achieved through the attention procedure in document presentation. A medical group of text is obtained via a classifier. The experimental examinations are performed in four medical text data sets, including two medical history datasets, two medical records.

Another study (Wang, Barnaghi, & Bargiela, 2009) has shown the concept of subject relationship and leveraging knowledge theory with the probabilistic topic models studied, two algorithms were proposed for learning terminological ontologies. Experiments were carried out with various model parameters, and the domain experts evaluated studied ontology statements. They had also compared the outcomes of our approach on the same dataset with two existing concept hierarchy learning methods. The analysis shows that in terms of recall and accuracy tests, our approach outperforms other techniques. The level of accuracy of the learned ontology is adequate for it to be deployed in digital libraries for browsing, navigation, and information search and retrieval purposes.

We found another study (Adhitama, Kusumaningrum, & Gernowo, 2017), To overcome the limitation of LDA labeling issues, the combined LDA and ontology. This study used databases for 50 news documents from an online news portal. The experiment found the best representation of word count for each topic to indicate the correct name of the subject. "The ontological method used is based on "Kamus Besar Bahasa Indonesian (KBBI)" field dictionary. Cohen's kappa's coefficient is used to calculate the reliability of the markings based on the approval of two language experts on the measurements on the label with a particular word representation which has more than 41% kappa value. The results of the analysis indicate that the highest kappa value of the five terms in each subject is 100 percent while the highest mean value is 80 percent in each subject's 5 words and 30 words.

A paper described a proposed NN classificatory of propagation (Govindarajan & Chandrasekaran, 2007) which executed cross-validations in the NN. Also, it solved the original NN classification problem. Less training time reduced the accuracy of classification expansion. The viability of the benefits of the proposed solution can be seen in 5 data sets including contact lenses, CPUs, symbolic temperatures, negative data about the weather, and labor. It is shown that when the data set is larger than the CPU or the network is equipped with many hidden units. Also, it reduced training time more than 10 times faster compared to the output of the NN model by CPUs, symbolic temperatures, negative data on weather, and labor. This happens when the data set is larger than the CPU or the network is equipped with many unknown units, which are the only missing attributes. On average, the accuracy of the proposed NN for touch lances was about 0.3 percent lower than that of the original NN. To ensure that there can be many ideas and solutions that can be transferred to various classification paradigms, this algorithm is independent of these data sets.

In LDA there are inefficiencies in automatically labeling each paper separately, so some ideas for automatic labeling have been seen in previous Recurrent Neural Network (RNN) research. Artificial Neural Network (ANN) (Singh et al., 2000) have been suggested, and The efficient results that are used in texting classification due to the natural syntax structure that is ideally suited to natural language processing are strongly achieved in recurrent NNs. But because of recurrent NN problems, for instance, when the length of the text series is too long the model is vulnerable to a gradient disappearing or a gradient explosion.

So we plan to come up with a full and full technology-based solution. Then the topic modeling, there were plenty of studies to automatically label but they mostly worked with a small number of texts, not like our study. So this is quite a challenge for us. After that in the LDA part, we found many inefficiencies while connecting with main topic classes. So we came up with prior studies on Ontology. The ontology

part of our research also gave trouble because in our case we got more than one label to a research paper. So we seek for automatic Classification Finally we got the ANN model to solve that problem.

"Current trends in wildlife research" editorial has revealed the foremost studies to our research. The research was released by Mateo et al. (2016), which uses a bibliometric approach to denote current trending topics and wildlife areas. They used the Scopus search engine they have selected all publications containing "wildlife" in the title of the text, abstract or keywords, or the newspaper name (source title in Scopus) from 1984 to 2013, in the Life Sciences, Health Sciences, Physical Sciences, or Social Sciences and Humanities. The result was 51,436 files. Those chapters finally offer an updated overview of the knowledge learned from wild animal science and a guide to identifying recent trends and the remaining gaps that must be resolved following our evolving environment's new requirements.

There are no similar researches used topic modeling for past research papers. For the wildlife in Sri Lanka, there are taxonomy-based researches found. There also we can find very lack of technological usage. In 2017 (Zhu et al., 2017), a hrLDA strategy, a data-driven hierarchical topic model for the extraction of terminological ontologies from heterogeneous external documents, was defined. We are creating a hierarchical topic model, hrLDA, which does not require one to line at every level of a topic tree in the subject number or set the lengths of the topic path from the root to the leaves. They introduced relationship extraction into topic modeling, leading to less perplexity, as well as a multiple topic path drawing technique, which is an enhancement over the hrLDA 's proposed simple topic path drawing process. In this section, the key problem we discuss is the generation of terminological ontologies in an unsupervised fashion. The basic hrLDA definition is as follows. When people create a text, they start with several topics being chosen. Then, for each topic, they choose some noun phrases as subjects. Next, they come up with relationship triplets for each topic to explain this subject or its relationships with other subjects. Finally, they connect the subject phrases and relation triplets to sentences via reasonable grammar. The main topic is normally described with the most important relation triplet.

Chowdhury and Zhu (2019) published a research paper on "Towards the ontology development for smart transportation infrastructure planning via topic modeling" that has shown Initiate the creation of an integrated ontology that can assist with long-term transportation infrastructure planning and decision-making by proposing a preliminary taxonomy in this field. To this end, 20 visionary documents released by government agencies on transportation planning were compiled and analyzed using subject modeling techniques. In particular, two methods of topic modeling were used: LDA and Non-negative Matrix Factorization (NMF) models to extract significant and evolving concepts related to the planning of transport infrastructure. A preliminary taxonomy of transport infrastructure planning was then developed and introduced, leveraging the significant and evolving concepts.

Another study on "Semi-Automatic Terminology Ontology Learning Based on Topic Modeling" proposed by Rani, Dhar, and Vyas (2017) is Topic modeling algorithms, namely LSI & SVD and Mr. LDA for Ontology Learning (OL), were explored. The research and the experimental outcome provide ample evidence of the efficacy of using OL's Mr. LDA subject modeling. In terms of creating richer subject-specific information and semantic retrieval, the experimental findings in the paper demonstrate the efficacy of the proposed method. The construction of terminology ontology is a preliminary step for the optimization of semantic queries (Topics and Words Detection) for information management (Rani et al., 2017).

Another study on "A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling" in 2017 (Allahyari, Pouriyeh, Kochut, & Arabnia, 2017) suggested a Probabilistic topic model, each document is defined as a multinomial distribution over subjects and each subject as a multinomial

distribution over terms, aimed at discovering latent subjects in text corporations. While humans can infer an appropriate mark for each subject by seeing representative phrases of the subject at the top, it does not extend to machines. Techniques for Automated Topic Labeling aim to solve the problem. The main purpose of topic labeling strategies is to assign studied topics to interpretable labels. Based on the knowledge-based topic model and graph-based topic labeling process, we established a topic labeling strategy, KB-LDA. The findings confirm the robustness and efficacy of the KB-LDA technique on various text selection datasets. A key point that increases the subject coherence compared to the traditional LDA model is the incorporation of ontological principles into our model. (Allahyaria, Pouriyeha, Kochuta, & Arabniaa, 2009).

Allahyari et al. (2017) published research on "OntoLDA: Compared to past work in this field, the accuracy of the topic of labeling can be improved by taking ontology concepts instead of words alone, which usually represent topics through groups of terms selected from topics. A topic model based on ontology for automatic topic labeling. Two outputs have been produced. These are subjects that integrate ontological concepts into a single system with subject models, where each word is represented as a multinomial distribution over concepts and a multinomial spread over terms and a method of subject marking based around the ontological meanings of the concepts contained in the discovered topics. The selection of the best subject labels depends on the semantic pated relation and their ontological classifications. The results of experiments carried out in two separate sets indicate that an efficient way of producing concrete lab concepts to the topics discovered is to integrate ontology concepts as additional and more rich characteristics between subjects and words and to describe subjects in terms of concepts (Allahyari et al., 2009).

In a paper on "KB-LDA: The benefits of generalizing pattern-based facts" object-verb tuples were exposed in text documents of a joint learning basis on hierarchy, relations, and facts" in the topic model system. This modular approach provides space for the inductive KB structure to be further evolved by putting additional constraints on corporate entities as additional models (Movshovitz-Attias & Cohen, 2015).

A paper on "Probabilistic Topic Models for Learning Terminological Ontologies" has shown the concept of subject relationship and leveraging knowledge theory with the probabilistic topic models studied, two algorithms were proposed for learning terminological ontologies. Experiments were carried out with various model parameters, and the domain experts evaluated studied ontology statements. We have also compared the outcomes of our approach on the same dataset with two existing concept hierarchy learning methods. The analysis shows that in terms of recall and accuracy tests, our approach outperforms other techniques. The level of accuracy of the learned ontology is adequate for it to be deployed in digital libraries for browsing, navigation, and information search and retrieval purposes (Wang et al., 2009).

A NN Classification Analysis on "Classifier Based Text Mining for Neural Network" has been using NN techniques to measure training time for five data sets such as contact lenses, CPU, weather symbol, temperature, labor-nega-data, a text mining classifier was developed for a small collection of text data set work. First, based on data classification in several data collections, they used our developed text mining algorithms, including text mining techniques. After that, to measure the training time for five data sets, they use the outgoing NN. Experimental results show that for all datasets, including contact lenses, the precision ('percent correct') was the same, which is the only one with missing attributes. Text Mining is about applying to an unstructured text called Knowledge Discovery in Text (KDT) or text data mining or text mining information discovery techniques. The key tasks are known to be the NN that solves classification problems, the training set, the test set, and the learning rate. This is the set of input/output

patterns that they used to train the network and used to determine the efficiency of the network, set the rate of change. This paper describes a proposed back propagation neural net classifier that performs cross-validation for the original NN.

In this research (de Mello, Senger, & Yang, 2005), they have developed a TFIDF matrix that will be used to learn from the NNs for the ANN and then categorize the documents into predefined categories that are implemented in the following documents and explained. Given that very few individuals have investigated the NNs and their role in the current system in terms of text categorization, they decided to implement and demonstrate how the text is easily categorized using ANN methods as easy and has many advantages compared to traditional methods.

The study on "Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise" (Hendrycks, Mazeika, Wilson, & Gimpel, 2018), enables major mark manipulation efficiency gains in robustness. Besides, the use of a clean label collection of precise data reduced particularly bad label noise. By proposing a loss correction technique that used specific examples in a data-efficient way to minimize the effects of label noise on deep NN classifiers, they use accurate data. They experimented with distinct label noises at several strengths through vision and natural language processing tasks and show that their approach greatly outperforms current methods. They studied the model characteristics needed to demonstrate and express such regularity in word vectors. The result was a new global logbilinear model of regression that combines the advantages of global matrix factorization and local context window methods of two main families of the literary model. By training on null elements only in a word-word correspondence matrix instead of on the entire sparse matrix or individual background windows, the model uses statistical information effectively.

Another study (Liu et al., 2019) has revealed on "Thirdhand smoke has been a growing topic for years in ChinaThird-hand smoke (THS)" consists of residual cigarette smoke emissions that linger on the surface and in the soil. These contaminants were either re-emitted as gas or react to produce secondary pollutants with oxidants and other environmental compounds. Collecting THS media reported from major media outlets and analyzing this subject using theme modeling would facilitate a deeper understanding of the role played by the media in transmitting this health issue to the public. Useful information can be provided by data analysis and visualization of news articles. Their study demonstrated that topic modeling can provide insight into the perception of THS-related news reports. This study of media trends has shown that the main concerns of the Chinese media reporting on THS are the associated diseases, air and particulate matter, and regulation and restrictions. More comprehensive reporting on THS based on scientific evidence and with less emphasis on sensational news also needs to be provided by the Chinese press. To verify and calculate the effect of THS-b, additional research related to sentiment analysis of news data is recommended.

Another study proposed a neural embedding approach to automatically label topics using Wikipedia titles. Their approach merged text and word embedding to select the most fitting labels for the topics. Compared to the state-of-the-art competitor method, their model was easier, more competitive, and produced better outcomes across a variety of domains. The subjects created by the theme models are usually presented as a list of words. To reduce the cognitive overhead of understanding these topics for end-users, they suggested the marking of a subject with a succinct word that summarizes the theme or meaning.

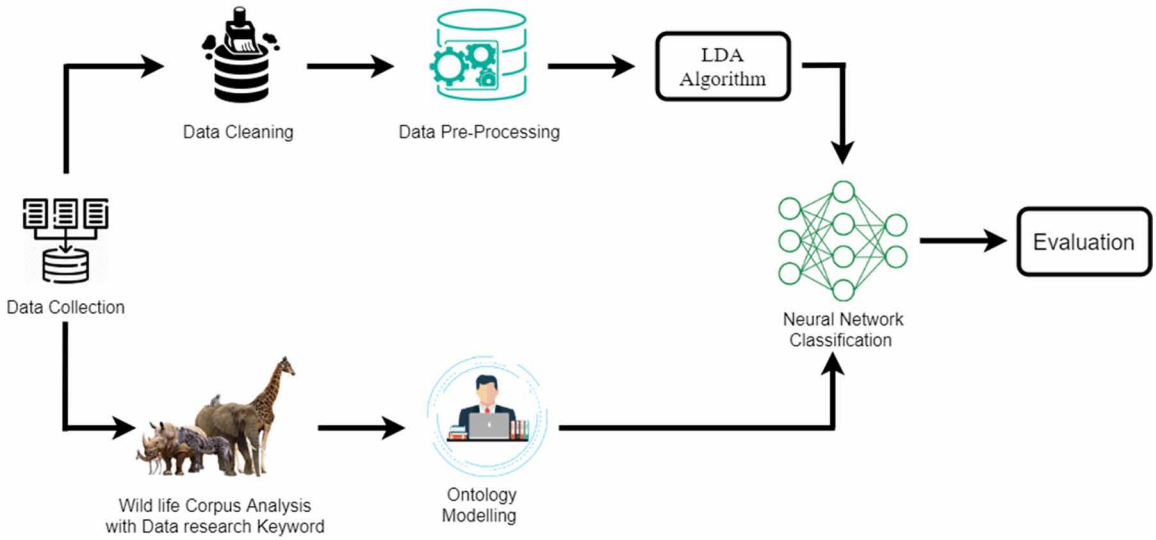
Finally, there were more studies on our related area. But we found their inefficiencies to overcome from our study. The especially first part of our literature review made by the key searches for wildlife in Sri Lanka we found a lot of lack in technology usage of this area. So we plan to come up with a full and full technology-based solution. Then the topic modeling, there were plenty of studies to automati-

cally label but they mostly worked with the small number of texts, not like our study. So this is quite a challenge for us. After that in the LDA part, we found many inefficiencies while connecting with main topic classes. So we came up with prior studies on Ontology. The ontology part of our research also gave trouble because in our case we got more than one label to a research paper. So, we seek for automatic Classification Finally we got the ANN model to solve that problem. So, in the conclusion of our literature review part, we worked for nearly 3 months with more than 500 papers to find the solution to our entire problem of this study.

Research Approach

We used integrated technologies in our methodology which shows in Figure 1. This methodology developed using ANN, LDA, and Ontology in this study. The text data of the defined domain were collected and pre-processed for the input to LDA algorithms then compared with the ontology graph to label the dataset the final output. After that ANN was used to classify. The steps of our methodology are defined below.

Figure 1. Methodological framework



Data Collection

We collected information about past wildlife researches in Sri Lanka from 2006 to 2019, with the aid of the Department of Natural Resources, Sabaragamuwa University of Sri Lanka, and an extreme literature survey. After that, we accessed full research papers of selected papers from each domain. We've selectively applied the title and abstract data to the CSV file from those research papers.

Data Cleaning

Data cleaning is the method of preparing data for review by deleting or altering data that is inaccurate, incomplete, obsolete, duplicated, or incorrectly formatted. Typically, this data is not necessary or helpful when it comes to analyzing the data because it can complicate the process or provide incorrect results.

We performed the following steps:

- Tokenization: Divide the text into sentences, and the sentences into words. Lower case the words and smooth punctuation
- Stop word removal: Delete all stop words. Natural Language Toolkit has a set of stop words modules.
- Lemmatizing: Words in the third person are shifted to first-person and verbs shifted to present from past and future tenses.
- Words are stemmed — words are reduced to their root form.

Data Preprocessing

Data pre-processing is so important because if our data set contained mistakes, redundancies, missing values, and inconsistencies that all compromised the integrity of the set, we need to fix all those issues for a more accurate outcome (Mateo et al., 2016). We used GloVe for the preprocessing works of our text data. Glove stands for Global Vector for Word representation which provides a high-level preprocessing vocabulary close to the pre-trained embedding (Brennan, Loan, Watson, Bhatt, & Bodkin, 2017). So we can get preprocessing to result in tokens that are mostly covered by word vectors.

Topic Modelling-LDA

LDA helped adapt the textual data into a format that could act as an input to the LDA model for training. We began by converting the documents to a simple representation of the vectors as a group of words called Bag of Words (BOW) (Rani et al., 2017). First, we translated a list of titles into vector lists, all with vocabulary-capable lengths.

Topic modeling is one of the unsupervised methods in trending. In other words, it is a text mining strategy in which is used to form topics for subjects or themes of documents that can be derived from a broader set of documents (Lee et al., 2018). LDA, one of the most famous & efficient modeling techniques, is a similarity model of a corpus based on Bayesian models. This is often considered a probabilistic extension of Latent Semantic Analysis (LSA). The LDA's basic idea is that each document has a word distribution that can be defined as.

Ontology Modelling

Ontologies contain features such as general vocabulary, reusability, machine-readable content, as well as ordering and structuring information for the Semantic Web application, enabling agent interaction, and semantic searching (Movshovitz-Attias & Cohen, 2015). Automated learning is the problem in ontology engineering, such as the lack of a fully automated approach to shape ontology using machine learning techniques from a text corpus or dataset of various topics.

The ontology model was finalized using protégé tools, which are the most popular tool of ontology visualization (Ahmed & Kovács, 2020). The Protégé 5.5.0 tool is being applied for further development in various disciplines for a better understanding of knowledge with the aid of domain professionals in the wildlife.

NN Classification

We used RNN classification to train and test the model of our automatic labeling process. Here to train our model, we used Long Short-Term Memory (LSTM). Three sections or layers, which are the input layer, the secret / intermediate layer, and the output layer, may characterize the NN. The input layer is used to receive the outer field's input signals. It is made up of neurons that go to the secret layer. The supervised-based method is used in NNs, so there is a response or output to the input provided to the NN. The NN processed input values and weights which took as input from the input layer and then goes to the hidden layer where the weights are summarized by the algorithm and the results are mapped to the proper output layer units.

For the training, we used 70% of our data set and testing 30% of the data. We input the abstract of each paper and train through ontology classes. We selectively trained for 7 major classes.

Evaluation

In our research, we used the output analysis method and compared our model with a manual classification which is used to assess the outcomes of the research concerning its objectives. Also, we used a web-based tool called OOPS is used to ontology model evaluation which allowed for the detection of anomalies in ontologies, independent from any ontology development environment. We detected anomalies with OOPS and solved those anomalies with the help of the domain specialist consultancy. The ontology includes a cycle between two hierarchy classes. It could avoid problems of modeling and reasoning to detect this situation (Lin, 2017). The NN prediction model was compared with manual classification to check the accuracy also the LSTM layers and the dataset allocation were increased according to get the maximum accuracy of 83% with manual classification from expert works.

RESULTS AND DISCUSSIONS

The results of this study were represented using abstract of past research papers which serves as an input in Sri Lanka. We used python language for LDA implementation. The text used as input is interpreted and tokenized with the result that input nouns, adjectives, and verbs are compiled. Also, it removes all the stop words in the research papers.

The tokenized and pruned text is then subjected to the algorithm of LDA modeling. That created word sets that could collect words that are connected as word sets. These word sets are listed as separate subjects. To organize, synthesize a large corpus, and retrieve subjects and words, the LDA model approach is used.

Figure 2 is the final visualizations of the LDA model which shows the overall keyword for each research paper and the essential keyword using the pyLDAvis library in python. This output allowed the detection of hidden keywords from every abstract. To get the output of the pyLDAvis method we used

the equation of saliency and relevance to accommodate the keyword distributions. The intertropical distance map is indicated via multidimensional scaling by our LDA output. In CE literature and inter-topic distance, the top 30 salient keywords.

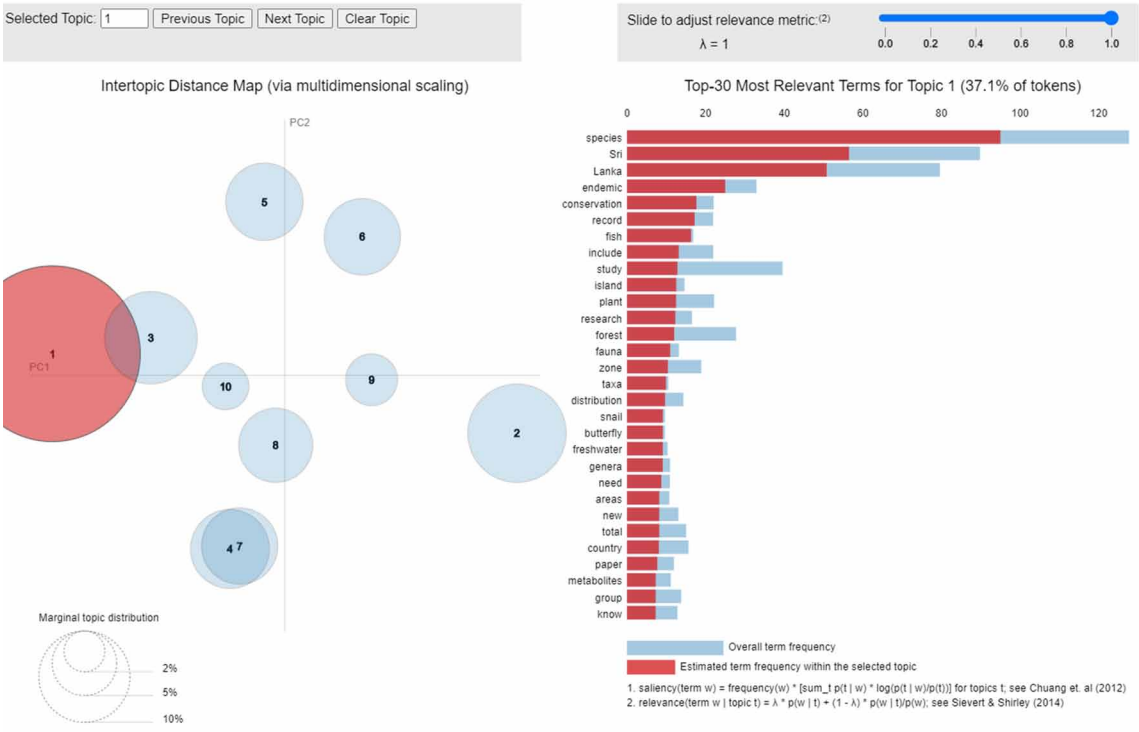
Saliency is used to describe the overall term frequency according to your data set. In our all research papers we found the top 30 overall terms to form a similarity cluster to form each topic. The saliency (Chuang, Manning, & Heer, 2012) is defined by the blue color bard in the given graph.

$$Saliency = frequency \times \left[sum\ p(t \mid w) \times \log\left(p\left((t \mid w) / p(t) \right) \right) \right] \quad (1)$$

Where (1), t- Topic, Frequency (w) –frequency of word w, p (t|w) - conditional probability: the probabilistic which identified word w was generated by latent topic t, p (t) - the probability of topic t, sum p (t|w) - the sum of the probability of identified word w was generated by latent topic t.

This formulation (1) defines (in a theoretical context of information Sense) the meaningfulness of specific term w, versus a randomly selected word, is for obtaining the generating subject. In case, if a word w appears in all topics, observing the word tells us nothing about the topical mixture of the document; thus the word will obtain a score of low distinctiveness.

Figure 2. LDA topic model output



Relevance is used to express the estimated term frequency compared to overall term frequency which means saliency in the given corpus. Relevance is used to measure the similarity between each keyword to find the final topics.

$$Relevance = \lambda * p(w|t) + (1-\lambda) * p(w|t) / p(w) \quad (2)$$

Where (2), λ –slide to adjust relevant metric, p (w|t) - conditional probability: the likelihood that identified word w was generated by latent topic t, p (w) –the probability of word w (Frasier et al., 2020).

The extracted major keywords of research papers were used as input to Ontology modeling. With the help of domain expert consultation, we modeled the ontograph which shows in Figure 3 to predict the suitable title of research papers. Using this output from LDA we compared the ontology output. Analyzed the estimated keywords and their ontology domain formation. The protégé tool used the Sri Lankan wildlife research domain ontology to be developed.

Figure 3. Ontograph final output

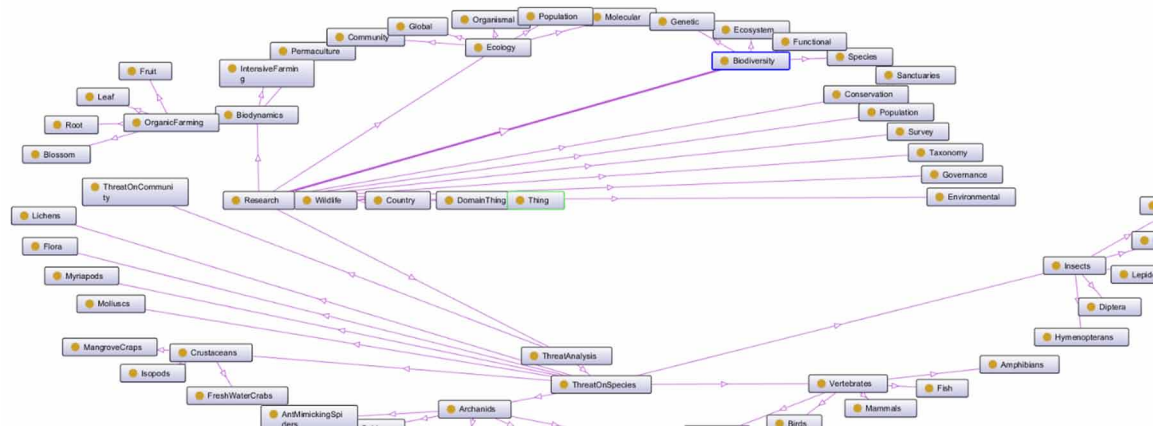


Table 1 shows the LDA output of some sample research papers.

Table 1. LDA output keyword extraction sample

Paper Id	Paper Name	Top 3 Main keywords	Main class
1	Characterization of Daboia russelii and Naja naja venom neutralizing ability of an undocumented indigenous medication in Sri Lanka (M. M. Silva, S. S. Seneviratne, D. K. Weerakoon, and C. L. Goonasekara,2017)	Venom, Herbal, Preparation	Reptile
3	Marine bacteria and fungi as sources for bioactive compounds: present status and future trends (V. Gunathilake,2012)	Marine, bacteria, fungi	Biodiversity
6	Reptile diversity in beraliya mukalana proposed forest reserve, Galle district, Sri Lanka (D. M. S. S. Karunarathna and A. A. T. Amarasinghe,2012)	Forest, reptile, anthropogenic	Reptile
10	Changes in soil carbon stocks under different agricultural management practices in North Sri Lanka (R. R. Ratnayake, T. Kugendren, and N. Gnanavelrajah,2014)	Soil, fertilizer, fraction	Permaculture

After the LDA keyword extraction and Ontology model output, we performed the classification manually with the help of domain expert advice.

For the automatic classification part, we need the labeled data set as the input of ANN. The model trained through 3 major layers as Embedding, LSTM, and dense layers (Shishov, 2017). LSTM is a kind of Recurrent ANN. It allows for the iteration of the model for efficient output. Also, the ontology model provided the best-labeled dataset to the ANN. Because learning a text like an abstract of a research paper is a little bit hard to implement. The dataset after cleaning used to vectorize because ANN only understands Float instead of text. Figure 4 has shown the vectored classification for each class. From ontology, we selected the 7 major classes. Figure 5 describes the accuracy of the ANN model through LSTM.

After that, we calculated the loss and accuracy to measure the deviations between predictions and real outputs. We used the Sri Lankan research papers for the testing of the model. In our research, the NN model's accuracy was 80% with the less and sensitive data we achieved this number.

Ontology specified the explicit classes of the wildlife domain of a given dataset through keywords of the research papers. Using the expert consultation we modeled the ontograph and with the help of LDA output, using ontology output we labeled the dataset for the ANN input. Then we vectored the data for a float from the text. Using ANN we trained our model for the specified classes for labeling the research papers. The model had 3 major layers as embedding layer, LSTM layer, and dense layer. Table 2 shows the feature's values to the layers of the ANN model. Figure 4 describes the classification vectors for each class.

Table 2. ANN feature value details

Features	Value
Epoch	100
Batch size	50
Optimizer	Adam
Loss	categorical_crossentropy
Activation Function (all dense layer)	ReLu
Activation Function (Final output layer)	SOFTMAX

Finally, the model tested using test data as research papers of wildlife in Sri Lanka. Figure 5 portrays the accuracy of the model.

The gap between training and validation shows the accuracy of the ANN model. For the manually labeled dataset, 83% accuracy was a very remarkable amount.

Table 3 shows the comparison of Major topics using Manual Classification and Major topics using LSTM Model Classification for some selected papers.

Figure 4. LSTM classification model

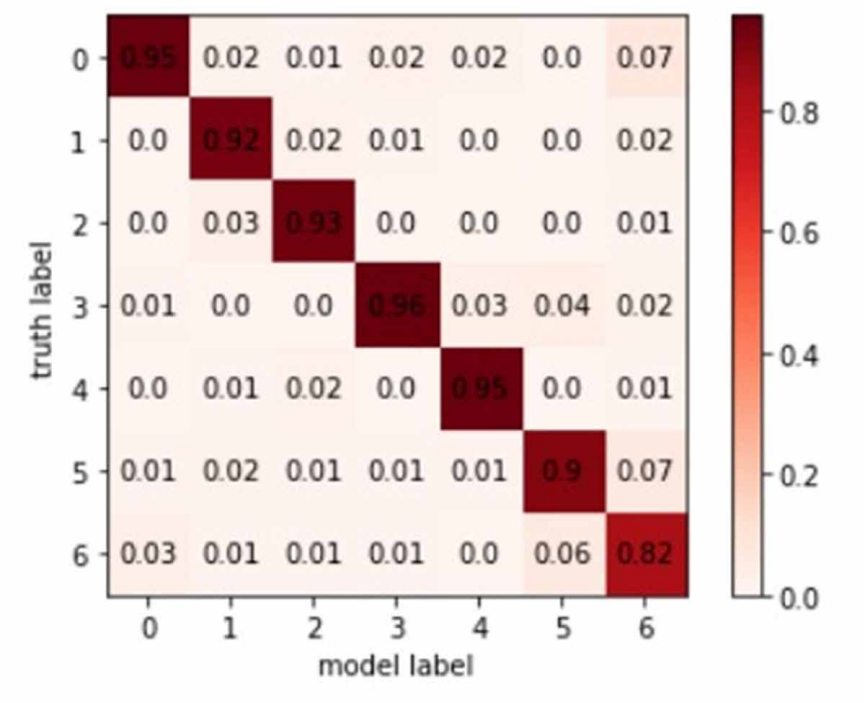


Figure 5. Accuracy graph for the LSTM model

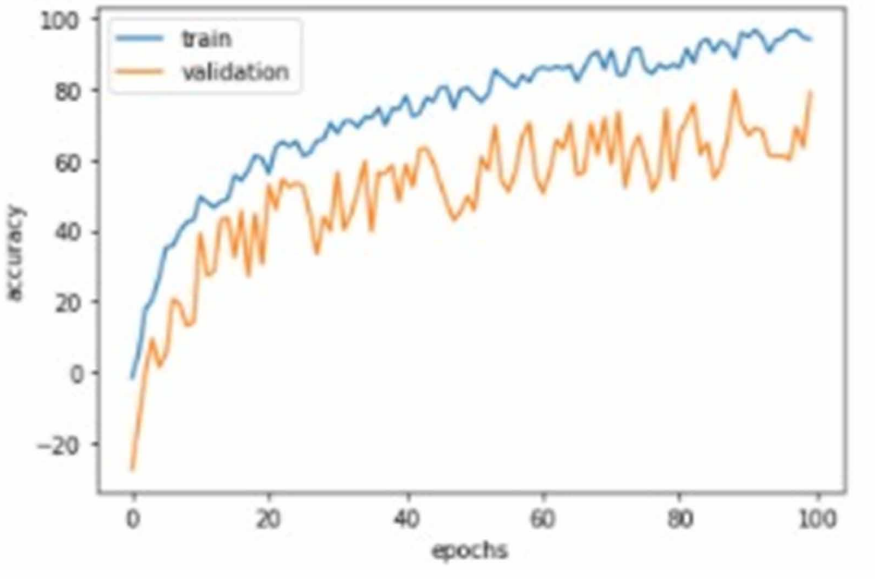


Table 3. Comparison of evaluation for classification

Paper Id	Paper Name	Major topic using Manual Classification	Major topic using LSTM Model Classification
1	Marine bacteria and fungi as sources for bioactive compounds: present status and future trends (V. Gunathilake,2012)	Microorganisms	Biodiversity
2	Reptile diversity in beraliya mukalana proposed forest reserve, Galle district, Sri Lanka (D. M. S. S. Karunarathna and A. A. T. Amarasinghe,2012)	Reptile	Reptile
3	Changes in soil carbon stocks under different agricultural management practices in North Sri Lanka (R. R. Ratnayake, T. Kugendren, and N. Gnanavelrajah,2014)	Permaculture	Permaculture
4	Characterization of Daboia russelii and Naja naja venom neutralizing ability of an undocumented indigenous medication in Sri Lanka (M. M. Silva, S. S. Seneviratne, D. K. Weerakoon, and C. L. Goonasekara,2017)	Snake	Reptile
5	An Overview of the Taxonomic Status of Class Hexapoda (Insecta) in Sri Lanka (Anura Wijesekara, 2015)	Insects	Insects
6	Current Status and Future Directions in Bee Taxonomy in Sri Lanka (W.A. Inoka P. Karunaratne and Jayanthi P. Edirisinghe,2014)	Bee	Bee
7	Current Status of Taxonomy, Research, and Conservation of Dragonfly Fauna (Insecta: Odonata) of Sri Lanka (Matjaz Bedjanic,2013)	Insects	Insects
8	Current Status of Aphid Taxonomy in Sri Lanka (Jayanthi P. Edirisinghe and M.A.P. Wijerathna,2011)	Insects	Insects
9	Current Taxonomic Status of Ants (Hymenoptera: Formicidae) in Sri Lanka (R. K. Sriyani Dias,2008)	Ant	Ant
10	Species Richness, Distribution, and Conservation Status of Butterflies in Sri Lanka (W. P. N. Perera and C.N.B. Bambaradeniya,2006)	Butterfly	Butterfly

Drawbacks

In this research, we tried to find out the areas in which the past research have been done in the context of Sri Lankan Wild Life. We used machine learning to narrow down the context and keywords mapped with the ontology were traversed to find the main keyword related to the research. But the main limitation is, that past research cannot be analyzed using a single keyword.

SOLUTIONS

If the past research context could be found with a few sets of keywords, then better results could be produced. Better results will be achieved if more datasets were focused to get more accuracy in this prediction model.

FUTURE RESEARCH DIRECTIONS

We plan to expand this research work to the worldwide research data. Also, in this research work we focused on the limited number of classes which is extracted from the ontograph . Therefore, if more

datasets were added, we can work for the whole specific classes of the ontology model. When other hybrid machine learning approaches can be utilized to extract the research theme of the past research then more predictions can be inferred.

CONCLUSION

In this paper, we have suggested an automatic labeling model for the research papers on the wildlife of Sri Lanka. We used a methodology which first LDA to extract the hidden keyword of each paper and after that Ontology to model the domain to label the data set with the help of domain experts. Using this method, the hidden keyword and the relations between the keywords also identify to help future research ideas. ANN is better than other text labeling algorithms based on our accuracy of the model. After the training of the test data, the model tests the testing data, and it shows the accuracy of every algorithm with different features. We manually compared the results from both LDA and terminology ontology. Both methods were effective but the RNN model gave better accuracy 83% than LDA and the ontology part 67%. When we worked with ontology and LDA we want accurate and deep knowledge about the domain but in RNN it's not necessary.

Finally, the dataset was trained and test through RNN. After finishing the model we got 83% accuracy which is relatively higher for text data that were domain-specific and comparatively higher amounts. We manually compared the results from both LDA-terminology ontology & ANN. Our study's output evaluation was 83% accurate to the overall conclusion. This work reduced the complexity to label the research papers without any domain pre-knowledge. The study focuses majorly on the limited classes of the ontology, for future research works will be work with more data set for each subclass of the ontology model, the final prediction model will be more wide range and accurate.

REFERENCES

Adhitama, R., Kusumaningrum, R., & Gernowo, R. (2017). *Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology.* Paper presented at the 2017 1st International Conference on Informatics and Computational Sciences (ICICoS). 10.1109/ICICOS.2017.8276370

Ahmed, G. H. A., & Kovács, L. (2020). *Ontology Domain Model for E-Tutoring System*. Journal OF Software Engineering & Intelligent Systems.

Allahyari, M., Pouriyeh, S., Kochut, K., & Arabnia, H. R. (2017). A knowledge-based topic modeling approach for automatic topic labeling. *International Journal of Advanced Computer Science and Applications*, *8*(9), 335. doi:10.14569/IJACSA.2017.080947

Allahyaria, M., Pouriyeha, S., Kochuta, K., & Arabniaa, H. R. (2009). *OntoLDA: An Ontology-based Topic Model for Automatic Topic Labeling*. IOS Press.

Brennan, P. M., Loan, J. J., Watson, N., Bhatt, P. M., & Bodkin, P. A. (2017). Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery. *British Journal of Neurosurgery*, *31*(6), 682–687. doi:10.1080/02688697.2017.1354122 PMID:28722516

Chowdhury, S., & Zhu, J. (2019). Towards the ontology development for smart transportation infrastructure planning via topic modeling. *Proceedings of the International Symposium on Automation and Robotics in Construction*.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the international working conference on advanced visual interfaces*. 10.1145/2254556.2254572

de Mello, R. F., Senger, L. J., & Yang, L. T. (2005). Automatic text classification using an artificial neural network. *High-Performance Computing in Science & Engineering*, 215–238.

Frasier, T. R., Petersen, S. D., Postma, L., Johnson, L., Heide-Jørgensen, M. P., & Ferguson, S. H. (2020). Abundance estimation from genetic mark-recapture data when not all sites are sampled: An example with the bowhead whale. *Global Ecology and Conservation*, *22*, e00903. doi:10.1016/j.gecco.2020.e00903

Govindarajan, M., & Chandrasekaran, R. (2007). Classifier-based text mining for the neural network. *Proceedings of World Academy of Science, Engineering and Technology*.

Hendrycks, D., Mazeika, M., Wilson, D., & Gimpel, K. (2018). *Using trusted data to train deep networks on labels corrupted by severe noise.* arXiv preprint arXiv:1802.05300.

Lee, J., Kang, J.-H., Jun, S., Lim, H., Jang, D., & Park, S. (2018). Ensemble modeling for sustainable technology transfer. *Sustainability*, *10*(7), 2278. doi:10.3390u10072278

Liu, Q., Chen, Q., Shen, J., Wu, H., Sun, Y., & Ming, W.-K. (2019). Data analysis and visualization of newspaper articles on thirdhand smoke: A topic modeling approach. *JMIR Medical Informatics*, *7*(1), e12414. doi:10.2196/12414 PMID:30694199

Mateo, R., Arroyo, B., & Garcia, J. T. (2016). *Current Trends in Wildlife Research* (Vol. 1). Springer. doi:10.1007/978-3-319-27912-1_6

Movshovitz-Attias, D., & Cohen, W. (2015). Kb-lda: Jointly learning a knowledge base of hierarchy, relations, and facts. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers). 10.3115/v1/P15-1140

Qing, L., Linhong, W., & Xuehai, D. (2019). A novel neural network-based method for medical text classification. *Future Internet*, *11*(12), 255. doi:10.3390/fi11120255

Rani, M., Dhar, A. K., & Vyas, O. (2017). Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, *63*, 108–125. doi:10.1016/j.engappai.2017.05.006

Shishov, B. (2017). *Mental Workload Estimation on Facial Video Using LSTM Network*. Paper presented at the ASRTU China-Russia International Conference on Intelligent Manufacturing, China.

Singh, S., Markou, M., & Haddon, J. (2000). Natural object classification using artificial neural networks. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. 10.1109/IJCNN.2000.861294

Wang, W., Barnaghi, P. M., & Bargiela, A. (2009). Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering*, *22*(7), 1028–1040. doi:10.1109/TKDE.2009.122

Zhu, X., Klabjan, D., & Bless, P. N. (2017). *Unsupervised terminological ontology learning based on hierarchical topic modeling*. Paper presented at the 2017 IEEE International Conference on Information Reuse and Integration (IRI). 10.1109/IRI.2017.18

KEY TERMS AND DEFINITIONS

Artificial Neural Network (ANN): An artificial neural network (ANN) is a piece of a computer system programmed to replicate the way the human brain analyzes and processes information. The foundation of artificial intelligence (AI) solves problems that, by human or mathematical criteria, would be impossible or complicated. ANNs have the capability of self-learning, meaning that more evidence is needed to obtain improved outcomes.

Conservation: Conservation is the conservation and preservation of these properties so that they can continue to survive for future generations. It involves preserving the diversity of animals, genomes, and habitats, as well as environmental functions, such as nutrient cycling.

Latent Dirichlet Allocation (LDA): LDA is a generative mathematical model that allows a series of results to be explained by unobserved classes that understand why certain parts of the data are identical. For example, if observations are words gathered in documents, it is argued that each text is a combination of a limited number of themes and that the appearance of each word is due to one of the themes of the document. LDA is an example of a theme model which belongs to the machine learning toolbox and, more generally, to the artificial intelligence toolbox.

Long Short-Term Memory (LSTM): LSTM networks are a form of a recurrent neural network capable of learning order dependency on sequence prediction problems. This behavior is required in complex problem domains such as machine translation, speech recognition, and more.

Ontology: Ontology includes the description, systematic naming, and specification of types, properties, and relationships between concepts, data, and entities that underpin one, several, or all spheres of discourse. More specifically, ontology is a means of showing the properties of the subject field and how they apply to it, by identifying a collection of definitions and categories that describe the subject.

Recurrent Neural Network (RNN): A recurrent neural network (RNN) is a type of artificial neural network that uses sequential data or time-series data. These deep learning algorithms are widely used for ordinal or temporal concerns, such as language translation, natural language processing (NLP), speech recognition, and image captioning; they are implemented into popular applications such as Siri, voice search, and Google Translate.

Research: Research is a comprehensive investigation process involving the compilation of data; the recording of critical information; and the review and evaluation of those data/information in conjunction with appropriate methodologies developed by particular technical fields and academic disciplines. Research is undertaken to assess the relevance of a theory or an interpretive framework; compile a body of substantive information and observations to share them acceptably, and produce questions for further inquiry.

Topic Modeling: Topic modeling is a tool for unsupervised document classification, analogous to clustering on numeric data, which identifies certain normal classes of things (topics) even though we're not sure what we're searching for. Topic modeling provides tools for arranging, interpreting, scanning, and summarizing broad electronic collections automatically.