Ontology Based Machine Learning Approach to Automatic Labelling for Research Papers on Wildlife of Sri Lanka

Premisha Premananthan
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
ppremisha@std.appsc.sab.ac.lk

Banujan Kuhaneswaran
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

Kumara BTGS
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kumara@appsc.sab.ac.lk

Enoka P Kudavidanage
Department of Natural Resources
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
enoka@appsc.sab.ac.lk

***Abstract* — Sri Lanka, being a global biodiversity hotspot, places great emphasis on biodiversity from an ecological perspective, socio-economic, and cultural factors. However, the wildlife of Sri Lanka is critically threatened due to several factors. Mainly human activities and needs supersede conservation measures. Lack of knowledge and technical support also hinder wildlife management activities. Findings of wildlife research studies could be incorporated into data-driven conservation and management decisions but the current contribution is not satisfactory. This research shows a novel data mining approach for finding hidden keywords and automatic labeling of past research work in this domain. We used the Latent Dirichlet Allocation (LDA) algorithms to model topics and identify the major keywords. Using the output of Topic Modelling an ontology model was also developed to represent the relationships between each keyword. Using the ontology instances we classified the research papers using Artificial Neural Network (ANN) to predict the labels for research papers in the wildlife domain. These approaches can be used for guiding future research endeavors, with the recognition of research gaps and by classifying the subjects related to a publication by the non-professional related fields. The experimental results demonstrated a 83% of accuracy for the proposed method.**

Keywords — ANN, LDA, ontology, topic modeling, wildlife

I. INTRODUCTION

Wildlife is critical for the sustenance of life on earth. Biodiversity conservation is crucial to preserving a stable global ecological balance. Sri Lanka is a global biodiversity hotspot consisting of a large variety of fauna and flora. It is one of the main sources of income generation through tourism and other means. The diversity of ecosystems is primarily due to its topographical and climatic heterogeneity, as well as its coastal effect [1]. This rich biodiversity is threatened due to unplanned land use, pollution, overexploitation, etc. Our research mainly focuses to resolve the inadequate application of wildlife research and technologies in the decision-making process.

From a technological perspective, there was prior work [2] [3] that has shown hierarchical relationship-based latent Dirichlet allocation (hrLDA), a data-driven model of hierarchical topics to acquire terminology ontology from a large number of amalgamate documents. Unlike traditional topic models, hrLDA relies on noun phrases instead of unigrams, deals with syntax and text structures, and enriches topic hierarchies with topic relations. Through a series of experiments, we are demonstrating hrLDA's superiority over established topic models, especially for hierarchy building.

In LDA there are inefficiencies to automatically label each paper separately and so prior research on Recurrent Neural Network (RNN) has shown some ideas for automatic labeling. They suggested convolutional neural networks [4], recurrent neural networks are widely used in text classification because of their natural sequence structure, which is suitable for natural language processing. However, there is a well-known problem with recurrent neural networks, that is, when the length of the text sequence is too long, the model is prone to gradient disappearance or gradient explosion.

Their approach incorporates text and word embedding to pick the most appropriate labels for the topics. Compared to the state-of-the-art competitor method, our model is easier, more effective, and produces better performance across a variety of domains

Our objective of the research is to provide technical solutions to find hidden keywords and research ideas using past research papers on wildlife in Sri Lanka. The main goal is to implement a model that can label given research papers automatically. Also using LDA and ontology model we have to find topics and the relations between each topic to improve research ides in a given circle.

This paper is organized as follows. In Section II, we define the core theories used by the proposed methodology, Section III, we discuss the results of the study. In Section IV, we discuss the conclusion of our experiment and suggest areas for future study at the end.

II. METHODOLOGY

We used a semi-automated methodology which shows in Fig 1. This methodology developed using 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 the final output.

A. 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.

ISSN 2756-9160 / November 2020.
International Conference on Advances in Computing and Technology (ICACT–2020) Proceedings

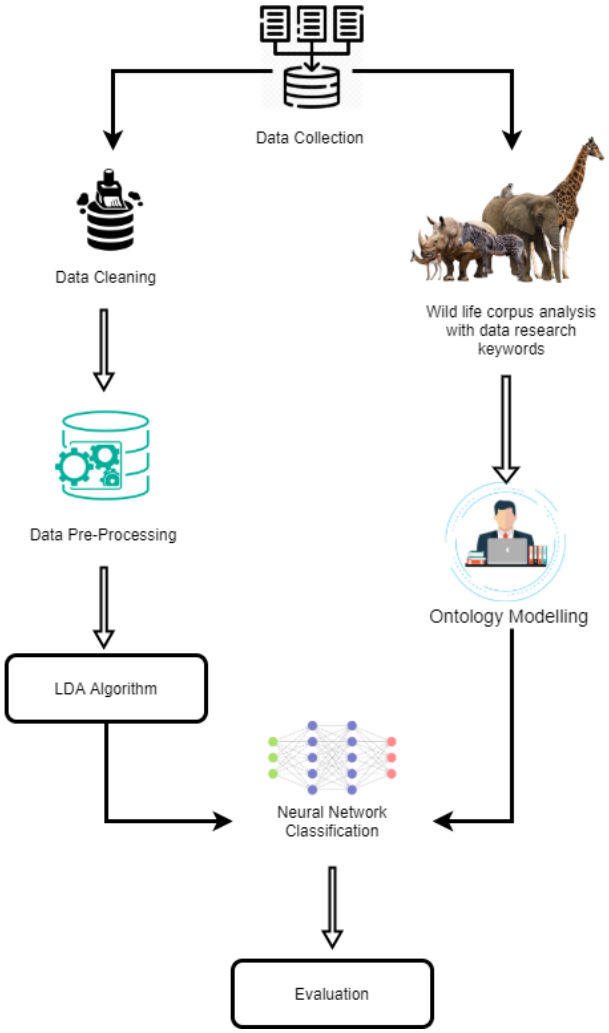

Fig. 1. Methodology

B. 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 words that have fewer than 3 letters. All stop words are removed.
- 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.

C. 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 [5]. 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 [6]. So we can get preprocessing to result in tokens that are mostly covered by word vectors.

D. 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). Using LDA we generated the major keywords and using those keywords, clustered the related keywords into specific topics. The topics from LDA were used as input to form the Ontology model.

E. 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 [7]. 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. So we used LDA to identify the major keywords. After using expert consultation we can easily model the Ontograph.

The ontology model was finalized using protégé tools, which is the most popular tool of ontology visualization [8]. 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.

F. Neural Network Classification

We used the RNN classification to train and test the model of our automatic labeling process. Here we used Long Short Term Memory (LSTM) to train our model. The neural network can be described in three sections or layers which are the input layer, the hidden / intermediate layer, and the output layer. The role of the input layer is to receive input signals from the outer field. It consists of neurons going to the secret layer. The learning of the neural network is completely supervised and thus the input given to the neural network has a response or output. The neural network takes input values and weights from the input layer as input and then goes to the hidden layer where the function sums up the weights and maps the results to the corresponding output layer units. ANN is the best suitable [9] for our model because the effectiveness of the model was high and our dataset was manually labeled using an ontology model and expert consultation.

III. Results and Discussion

The results of this study were represented using abstracts of past researches which serves as an input. We used python language for LDA implementation and ANN. The tokenized and pruned text is then subjected to the LDA modeling algorithm. That gave production as word sets that collection could contain words that are linked to each other. The hidden keywords from the data set were identified and similar keywords cluster as one topic for each research paper using the LDA visualization model as well as the inter topic distance also calculated.

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 I shows the feature's values to the layers of the ANN model. Fig 2 describes the classification vectors for each class.

Table 1 - 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

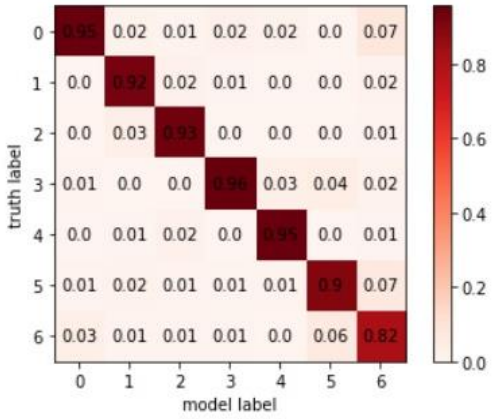

Fig. 2. ANN classification model

Finally, the model tested using test data as research papers of wildlife in Sri Lanka. Fig 3 portrays the accuracy of the model.

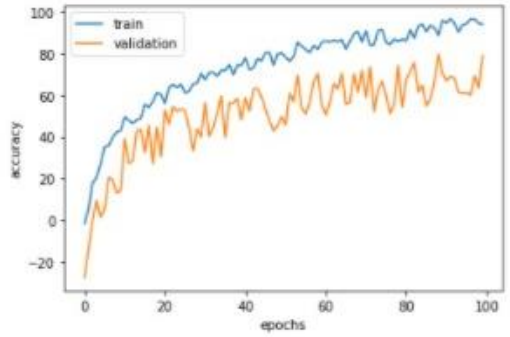

Fig. 3. Accuracy graph for the ANN model

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

IV. 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 uses LDA to extract the hidden keywords of each paper. Subsequently, ontology was developed to model the domain to label the data set with the help of domain experts. Finally, the dataset was trained and tested through RNN. After finishing the model, an accuracy of 83% was achieved, which is relatively high for domain-specific text data which. The results from both LDA-terminology ontology & ANN were manually analyzed.

This work reduced the complexity to label the research papers without any domain pre-knowledge. Using this method, the hidden keywords and the relations between the keywords are also identified to help future research ideas. It was observed that, ANN is better than other text labeling algorithms based on the accuracy of the model.

ACKNOWLEDGEMENT

We acknowledge the Department of Wildlife Conservation of Sri Lanka for the research permit (WL/3/2/60/15) granted to E. P Kudavidanage.

REFERENCES

[1] L.P.Jayatissa, *Present Status of Mangroves in Sri Lanka*. 2012.

[2] X. Zhu, D. Klabjan, and P. N. Bless, "Unsupervised terminological ontology learning based on hierarchical topic modeling," *Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017*, vol. 2017-Janua, pp. 32–41, 2017, doi: 10.1109/IRI.2017.18.

[3] S. Chowdhury and J. Zhu, "Towards the ontology development for smart transportation infrastructure planning via topic modeling," *Proc. 36th Int. Symp. Autom. Robot. Constr. ISARC 2019*, no. Isarc, pp. 507–514, 2019, doi: 10.22260/isarc2019/0068.

[4] S. Bhatia, J. H. Lau, and T. Baldwin, "Automatic labelling of topics with neural embeddings," *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, no. 1, pp. 953–963, 2016.

[5] N. T. R. Editors, *Technologies in Data Science and Communication*. 2019.

[6] P. M. Brennan, J. J. M. Loan, N. Watson, P. M. Bhatt, and P. A. Bodkin, "Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery," *Br. J. Neurosurg.*, vol. 31, no. 6, pp. 682–687, 2017, doi: 10.1080/02688697.2017.1354122.

[7] D. Movshovitz-Attias and W. W. Cohen, "KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 1, pp. 1449–1459, 2015, doi: 10.3115/v1/p15-1140.

[8] G. Hussein, A. L. I. Ahmed, L. Kovács, G. Hussein, and A. Ahmed, "ONTOLOGY DOMAIN MODEL FOR E-TUTORING SYSTEM," vol. 5, no. 1, pp. 37–44, 2020.

[9] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Nips, pp. 10456–10465, 2018.