

Snake species identification by using natural language processing

Nur Liyana Izzati Rusli¹, Amiza Amir², Nik Adilah Hanin Zahri³, R. Badlishah Ahmad⁴

^{1,2,3}School of Computer and Communication Engineering, Universiti Malaysia Perlis, Malaysia

⁴Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), 22200 Besut, Terengganu

Article Info

Article history:

Received Oct 1, 2018

Revised Dec 10, 2018

Accepted Dec 25, 2018

Keywords:

Natural language

Human perception

Snake images

TF-IDF

ABSTRACT

The paper presents the snake species identification by using natural language processing. It aims to help medical professionals in predicting the snake species for snake-bite treatments based on the patient's description of the snake. The decision in suitable anti-venom critically depends on the type of snake species. Wrong anti-venom may result in severe morbidity and mortality. This research investigates the human perception and the selection of words in describing a snake based on their visual view. The descriptions were presented in unstructured text, and the NLP processing involves pre-processing, feature extraction and classification. Four machine learning algorithms (naïve Bayes, k-Nearest Neighbour, Support Vector Machine, and Decision Trees J48) were used during training and classification. Our results show that J48 algorithm obtained the highest classification accuracy of 71.6% correct prediction for the NLP-Snake data set with high precision and recall.

Copyright © 2019 Institute of Advanced Engineering and Science.

All rights reserved.

Corresponding Author:

Amiza Amir,

School of Computer and Communication Engineering,

Pauh Putra Campus, Universiti Malaysia Perlis,

02600 Arau, Malaysia

Email: amizaamir@unimap.edu.my

1. INTRODUCTION

Snakes that are cold-blooded vertebrates falls into two categories; venomous and non-venomous. Many venomous snakes have appeared in many countries, and they are a real threat to the public safety and health. There are more than 3000 species of the snake nowadays, 600 of them are venomous, and over 200 are considered important in medical record [1]. Highest medical important treatments are snake bites from a highly venomous snake that are necessary to be recognized since they can cause severe pain and even death (e.g., Black Mamba, King Cobra, Indian Krait). Secondary medical important snake bites are due to the venomous snakes (e.g., Albino Burmese Python, Ball Python, Red Rat Snake) that can result in disability and severe pain but in less impacted due to their activity or maybe because of the habitat that near of human population. In Malaysia, the five-year review of snakebite patients shows that there were 260 cases of snake bites reported, and 52.9% of the snake bites were from unknown [2].

In many emergency cases, one has to identify the snake species merely based on the text description given to them by the victim or witness without any graphical aids. Being able to recognize the type of snake based on the description of the people have become very important in social and medical progression. To perform optimal clinical treatment, the diagnosis of the snake species responsible for the snake bite is crucial. The slightest delay might give result in severe morbidity and mortality. Thus, it is imperative to precisely and concisely determine the type or species of the snakes. The collected information is important to identify if the snake is venomous or not, thus helps medical professional to determine the suitable anti-venom and further treatment plan.

Typically, snake species are recognized manually based on the visual features such as head shape, skin color, eye shape, and body shape. This process requires knowledge of characteristics of the snakes which is not quite common for most people where only the experts have this useful knowledge. Considering the difficulty faced by most people in identifying the snake species, the main aim of this work is to perform species recognition based on the description from the witness or victim in unstructured text form.

In this work, an intelligent system that will help the medical professional to be able to predict the type of the snake based on the description in the unstructured text by using natural language processing (NLP). Common perception and words used by many different people to describe many different types of snakes will be analyzed. The text will be preprocessed, relevant keywords will be extracted based on their weight in the context, and these keywords will be used as features during classification by machine learning to learn and predict the species of snakes.

Limited studies have been conducted for species recognition by using machine learning. Butterfly species recognition in [3] uses neural networks to recognize butterfly species based on butterflies' shape. The branch length similarity (BLS) entropies from the boundary pixels of a butterfly shape were extracted in this study. Wood species recognition was proposed by Zhao et al. [4, 5] and Zamri et al. [6]. In [4] and [6], image based features (color, texture, and spectral features) were extracted to identify the wood species by using the back propagation neural network. In [5], k-nearest neighbor (k-NN) was used to classify wood species through images. Image-based plant species recognition by using k-NN was also suggested by Faria et al. [7].

Christiansen et al. [8] use a k-NN classifier to discriminate animal and non-animal based on heat characteristics of objects. While in the work of Yu et al. [9], Support Vector Machine (SVM) has been used to extract features and classify images of 57 animal species captured by camera traps with an average classification accuracy of 82%.

To our knowledge, the closest work to our research can be found in [10] and [11]. In these works, automatic snake species identification techniques from snake images were proposed by using machine learning algorithms. Amir et al. [10] applied texture based approach as features, while James et al. [11] used features describing top, side and body views of snake images. In contrast, NLP was utilized in our work to enable snake species recognition through text-based information from a human.

2. RESEARCH METHOD

This work involved the collection of the text-based description of snakes based on the presented snake images by using survey methods (questionnaire). Then, important features were extracted by using term frequency – inverse document frequency (TF-IDF), and these features were provided to machine learning algorithms to learn and predict the snake species using Weka tool [12].

2.1. Raw Data Collection

60 respondents from multiple ranges of age participated in the questionnaire survey during data collection process. The respondents were shown with series of snake's pictures. Images of three species of snakes (Naja Tripudians, Boa Constrictor, and Dog-Toothed Cat) were used in this survey. After that, the respondents were asked a few questions in a questionnaire to describe the snake image that they had seen based on their perception and opinion of the snake. They were allowed to use their own words to explain the snakes' characteristics. Two examples of snake images are shown in Figure 1.

During the survey, the respondents are guided towards explaining the eight physical characteristics of the snake – that plays a major role in deciding what kind of the snake that is venomous or non-venomous. A snake observer is always using these characteristics to recognize the snake species:

- a) Length of its body
- b) The shape of its body
- c) Its head and neck shape
- d) The color and pattern on its body
- e) Scale texture
- f) Eye pupil shape
- g) Tail scales
- h) Anal plat division

180 samples of unstructured text represent the snakes' description are obtained from the questionnaire.



Figure 1. Two examples of snake images from the species of Dog-Toothed Cat (a) and Boa Constrictor (b)

2.2. Text Pre-processing

The method of pre-processing text is the first step and an important step in text mining techniques. Pre-processing is performed to minimise the dimensionality of the representation space which included [13]:

a) Data tokenization

Tokenization was performed using Weka to break down a text into pieces of words. In this work, tokenization is broken into words. Example of tokenization is shown as follows:

Input: I saw a green snake, and it has two fangs

Output: I, saw, a, green, snake, and, it, has, two, fangs

b) Stemming

Stemming is the process of finding the root of the word from different word forms, where the suffixes and prefixes will be removed. A word such as “playing” and “played” can be stemmed as “play”. Stemming was needed as it prevents overflow of the different word with the same meaning in the libraries. Example of stemming process shown as follows:

Input: I, saw, a, green, snake, and, it, has, two, fangs

Output: I, see, a, green, snake, and, it, has, two, fang

c) Symbols and Stop-word elimination

The stemmed text obtained previously underwent the process in removing all the special symbols such as ‘(’, ‘)’, ‘#’, ‘!’, ‘?’, ‘_’, ‘+’, ‘-’, ‘*’, and ‘/’. Stop word or stop word list are the set of common word that human use every day in any language. It does not have less significant meaning in the text or paragraph. Common words (e.g. “a”, “an”, and “the”) are eliminated by using stop-word removal function in Weka. This process can minimise the dimensionality about 15% to 20% reduction in the collected data [13].

Input: i, see, a, green, snake, and, it, has, two, fang

Output: i, see, green, snake, it, has, two, fang

2.3. Feature Extraction using TF-IDF

A high number of words in the text description will cause high dimensionality of the representation space during training and testing. Therefore, in this work, TF-IDF weighting is used to identify important and relevant keywords from each description. These high weighting keywords will be extracted as important features and will be used to optimise the training process. Term frequency (TF) represents how many times the number of the word that occurs in a single text or document. We used a flexible filter named StringtoWordVector in Weka to convert string attributes into a set of word vector which represents the words occurrence. Below are three examples of a snake description by a human in text form and how feature extraction is done.

Example:

Text 1: “I saw a long and a green snake.”

Text 2: “The green snake is a dangerous snake.”

Text 3: “The long snake is scarier.”

After pre-processing, the text will be as follows:

Text 1: I, see, long, green, and, snake

Text 2: green, snake, be, dangerous, snake

Text 3: long, snake, be, scary

TF-IDF weight of a term is calculated as follows:

(a) Calculate term frequency (TF)

(b) Calculate document frequency (DF) and the inverse of the DF (IDF).

(c) Compute TF-IDF

The normalized TF is measured according to Equation (1).

$$\text{Normalized TF} = \frac{\text{No. of term that occurred in the text}}{\text{Total no of word in the text}} \quad (1)$$

In reality, each text will contain different size, and usually, the value of TF will be higher than an example of TF in Table 1. Next, the text will be normalized based on its size by dividing TF by the total number of terms.

Table 1. TF Basic Calculation for Text 1, Text 2 and Text 3

Text 1	i	see	long	and	green	snake
TF	1	1	1	1	1	1
Norm TF	0.167	0.167	0.167	0.167	0.167	0.167
Text 2	green	snake	be	dangerous		
TF	1	2	1	1		
Norm TF	0.200	0.400	0.200	0.200		
Text 3	long	snake	be	scary		
TF	1	1	1	1		
Norm TF	0.250	0.250	0.250	0.250		

In TF, all terms being treated as equal. In contrast, the inverse document frequency (IDF) is a measure of how much information the word provides across all text or documents. Thus, IDF is computed as following Equation 2:

$$IDF = \log\left(\frac{\text{Total number of texts}}{\text{No. of text in which selected term is appeared}}\right) \quad (2)$$

For example, the term of the “green” was used to find IDF:

Total no. of texts: 3

Number of texts with term green on it: 2

$$IDF(\text{green}) = \log\left(\frac{3}{2}\right)$$

Table 2 shows the example of measured IDF value for terms that appeared in all the text. Finally, TF-IDF weight is measured using Equation (3).

Table 2. Inverse Document Frequency

Terms	IDF
I	0.477
see	0.477
long	0.477
and	0.477
green	0.176
snake	0.000
be	0.176
dangerous	0.477
scary	0.477

$$TF_IDF = \text{Normalized TF} * IDF \quad (3)$$

Table 3. Example on word occurrences in TF-IDF

Words	Text 1	Text 2	Text 3
i	0.080	-	-
see	0.080	-	-
long	0.080	-	0.120
and	0.080	-	-
green	0.029	0.035	-
snake	0.000	0.000	0.000
be	-	0.352	0.044
dangerous	-	0.095	-
scary	-	-	0.120

From the example in Table 3, the word “snake” is considered common due to it is an occurrence in all description with weight value of 0.000. In another word, the word “snake” is not significant in determining the characteristic of a snake described in each text. Therefore, a word with zero or low weight will be considered irrelevant feature and omitted during training and classification.

2.4. Training and Classification

In supervised classification, training must be first conducted, and classification task follows this. The training involves building a model based on one or more numerical and categorical variables such as attributes or features. Classification is a text mining task of predicting the value of a categorical variable such as target or class.

Four machine learning algorithms were chosen to perform these tasks. They are naïve Bayes [13], Support Vector Machine (SVM) [14], k-Nearest Neighbours (k-NN) [15], and decision tree J48 [16]. In this work, 180 samples of text-based description collected from 60 respondents will be used for training and classification. Due to limited sample, in order to obtained more accurate result, 10-fold stratified cross validation was applied to ensure the validity of our result be conducted for each algorithm. The training and classification processes were performed on different sets of the data as to generalize the new information.

3. RESULTS AND ANALYSIS

The questionnaire was filled by the society through social media as the medium, and we obtained 180 samples from 60 participants. Each participant was asked to describe three snake images (representing a species each). The raw dataset in text form then was imported into an Attribute-Relation File Format (ARFF) file. Then, preprocessing and feature extraction were performed. ARFF file is less memory intensive, faster and better for analysis because it includes meta data about column header and data column.

Converting a word to a vector is simply a mechanism to input and process words for any natural language processing task. As mentioned in Section V, during preprocessing, Weka package was used to convert word into the vector. This results in 483 attributes were found in the dataset. Then, features extraction methods such as stop word elimination, stemmer, and tokenizer were performed. Finally, TF-IDFT transform was executed to calculate the weight of each word in each document.

These feature extraction tasks result in a reduction of the dimensionality of attributes to 30%. After features extraction, the number of attributes decreases to 346 attributes. Hence, the resulting data set which called the NLP-Snake data set consists of 180 samples with 346 attributes.

3.1. Classification Accuracy

Classification accuracy is presented as a percentage where 100% is the best an algorithm can achieve. Four machine learning algorithms decision tree J48, SVM (Linear Kernel), naïve Bayes, k-NN are selected as classifiers in this project. The performance of the classifiers as reported in Figure 2 illustrates the correctly and incorrectly classified instances. Figure 2 indicates that J48 has the highest percentage of 71.67% followed by SVM with 68.33%. Naïve Bayes obtained 61.11% which then followed by k-NN by 55.56% as the lowest percentage obtained for correctly classified instances.

For incorrectly classified instances, J48 obtained the lowest percentage of being incorrectly classified instances with 28.33%. This then followed by 31.67% for SVM, 38.89% for naïve Bayes. k-NN obtained the highest proportion of incorrectly classified instances by 44.44%. Hence, J48 achieves the highest percentage of correct prediction compared to SVM, k-NN, and naïve Bayes for the NLP-Snake dataset.

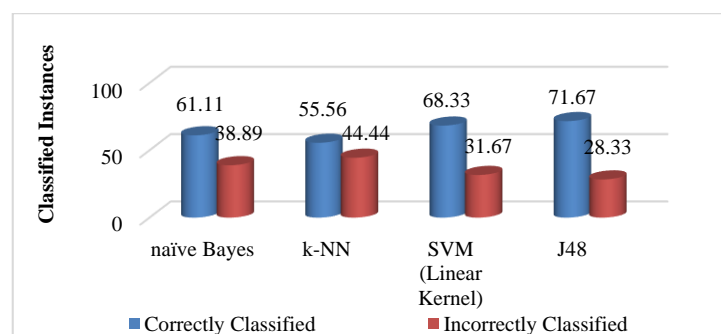


Figure 2. The accuracy of naïve Bayes, k-NN, SVM, and J48 for NLP-Snake dataset

3.2. Precision and Recall

Regarding probabilistic interpretation, precision and recall are not interpreted as ratios. Instead, they are interpreted as probabilities. Precision is the probability that a selected data is relevant while recall is the probability that a selected data is correctly retrieved. Precision and recall both are statistical measures of performances of a machine learning algorithm. The outcomes were shown in Figure 3 and Figure 4.

Figure 3 illustrates the precision performances obtained in Weka interfaces after ten training and classification were carried out. It shows that the highest precision outcome of machine learning algorithms for Boa Constrictor was k-NN by 87.2%, the second highest is J48 by 78.9%, the second lowest precision for Boa Constrictor is 62.1% followed by naïve Bayes as the lowest precision by 60%.

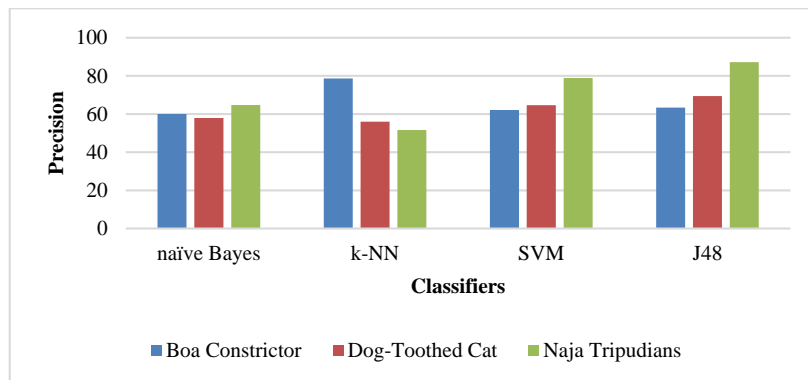


Figure 3. The precision performance of naïve Bayes, k-NN, SVM, and J48 for three snake species in NLP-Snake dataset

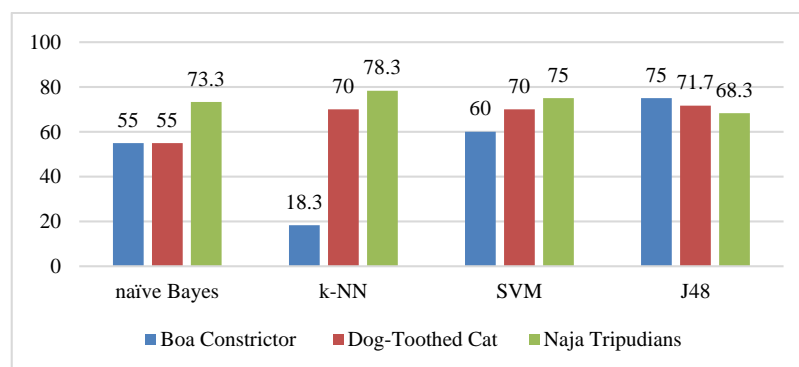


Figure 4. The recall performance of naïve Bayes, k-NN, SVM, and J48 for three snake species in NLP-Snake dataset

The highest precision of machine learning algorithms for Dog-Toothed Cat is J48 with 69.4%, SVM with 64.6%, followed by naïve Bayes with 57.9% as the second lowest precision and k-NN with 56% as the lowest precision. For Naja Tripudians, the highest precision is J48 by 87.2%; the second highest precision is SVM by 78.9%, followed by the second lowest precision for Naja Tripudians is naïve Bayes by 64.7%. Moreover, the lowest precision is obtained by k-NN with 51.6% only.

Figure 4 shows that the highest recall outcome of machine learning algorithms for Boa Constrictor was J48 by 75%, the second highest is SVM by 60%, the second lowest recall for Boa Constrictor is 55% in naïve Bayes followed by k-NN as the lowest recall by 18.3%.

The highest recall of machine learning algorithms for Dog-Toothed Cat is J48 that goes by 71.7%, SVM and k-NN share the same outcome by 70% and naïve Bayes by 55% as the lowest recall. For Naja Tripudians, surprisingly the highest recall is k-NN by 78.3%, the second highest recall is SVM by 75%, followed by the second lowest recall for Naja Tripudians is naïve Bayes by 73.3%. J48 obtains the lowest recall with 68.3% only.

In general, for precision and recall performance among the four algorithms for the NLP-Snake data set, J48 shows the most precise percentage were obtained for each snake species. It also has the highest accuracy compared to SVM, k-NN and naïve Bayes.

4. CONCLUSION

The paper demonstrates the preliminary result for the recognition of snake species by using natural language processing. Human description of snake images from three species was collected through a survey in social media. The resulting raw data set contains a textual description of snake characteristics by a human. The pre-processing and feature selection was then performed. The feature extraction task involves stop word elimination, stemming, word tokenizer and TF-IDF transform. The processed data set, named NLP-Snake dataset, consists of 346 attributes with 180 samples. Then, the performance of four machine learning algorithms (naïve Bayes, k-NN, SVM and decision tree J48) are evaluated for training and classification. All in all, the overall performances show that the J48 is the best and suited for text classification task, in particular, to identify snake characteristic in natural language task.

In the future, we aim to collect a larger data set by involving a greater number of snake species and more participants. By doing so, more accurate results are expected for real-world implementation.





ACKNOWLEDGEMENTS

We would like to thanks to the random participants who helped us by answering the survey for this study. We also would like to thanks Taman Ular Perlis for the snake pictures.

REFERENCES

- [1] WHO blood products and related Biologicals animal sera Antivenoms frames page. Retrieved October 7, 2016, from <http://apps.who.int/bloodproducts/snakeantivenoms/database/>
- [2] Chew, K.S., Khor, H.W., Ahmad, R., Rahman, N.A.H.N (2011). A Five-year retrospective review of snakebite patients admitted to a tertiary university hospital in Malaysia. *International Journal of Emergency Medicine* 4(1), 1-6
- [3] Kang, S.H., Song, S.H., Lee, S.H. (2012). Identification of butterfly species with a single neural network system. *Journal of Asia-Pacific Entomology* 15(3), 431 – 435.
- [4] Zhao, P., Dou, G., Chen, G.S. (2014) Wood Species identification using feature-level fusion scheme. *Optik - International Journal for Light and Electron Optics* 125(3), 1144 – 1148.
- [5] Zhao, P., Dou, G., Chen, G.S. (2014). Wood species identification using improved active shape model. *Optik - International Journal for Light and Electron Optics* 125(18), 5212 – 5217
- [6] Zamri, M.I.P., Cordova, F., Khairuddin, A.S.M., Mokhtar, N., Yusof, R. (2016) Tree species classification based on image analysis using improved-basic grey level aura matrix. *Computers and Electronics in Agriculture* 124, 227 – 233.
- [7] Faria, F.A., Almeida, J., Alberton, B., Morellato, L.P.C., Rocha, A., da S. Torres, R.(2015) Time series-based classifier fusion for fine-grained plant species recognition. *Pattern Recognition Letters* pp.
- [8] Christiansen, P., Steen, K.A., Jrgensen, R.N., Karstoft, H.(2014) Automated detection and recognition of wildlife using thermal cameras. *Sensors* 14(8), 13778
- [9] Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T. (2013). Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing* 2013(1), 1–10
- [10] Amir A., Zahri N.A.H., Yaakob N., Ahmad R.B. (2017) Image Classification for Snake Species Using Machine Learning Techniques. In: Phon-Amnuaisuk S., Au T.W., Omar S. (eds) *Computational Intelligence in Information Systems*. CIIS 2016. *Advances in Intelligent Systems and Computing*, vol 532. Springer, Cham
- [11] James, A.P., Mathews, B., Sugathan, S., Raveendran, D.K. (2014) Discriminative histogram taxonomy features for snake species identification. *Human-centric Computing and Information Sciences* 4(1), 1–11
- [12] M Imambi, S., & Sudha (2011). Pre-Processing of medical documents and reducing Dimensionality. *Advanced Computing: An International Journal*, 2(5), 15–24
- [13] Zaidi, N. A., Petitjean, F., & Webb, G. I. (2016). Preconditioning an Artificial Neural Network Using Naive Bayes. *Proceedings of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2016*, pp. 341-353
- [14] Cao J, Fang Z, Qu G, Sun H, Zhang D (2017). An accurate traffic classification model based on support vector machines. *Int J Network Mgmt.* 2017;27:e1962.
- [15] P. Gu, R. Khatoun, Y. Begriche and A. Serhrouchni. (2017). k-Nearest Neighbours classification based Sybil attack detection in Vehicular Networks. *2017 Third International Conference on Mobile and Secure Services (MobiSecServ)*, Miami Beach, FL, 2017, pp. 1-6.
- [16] M. Aashkaar, P. Sharma and N. Garg. (2017). Performance analysis using J48 decision tree for Indian corporate world. *2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS)*, Bangalore, 2016, pp. 1-5.

BIOGRAPHIES OF AUTHORS

	<p>Nur Liyana Rusli graduated with Bachelor of Computer Network Engineering from University Malaysia Perlis in 2017. She is now working in Ericson as a NOC Engineer. Her interest include machine learning and computer networks.</p>
	<p>Amiza Amir is a senior lecturer in School of Computer and Communication Engineering at Universiti Malaysia Perlis. She received her Ph.D. in Information Technology, on distributed artificial intelligence, from Monash University, Australia in 2015. Her current research interests include machine learning, distributed system, meta heuristic optimization, data analytics and software-defined network (SDN). She teaches courses in data analytics and artificial intelligence.</p>
	<p>NIK ADILAH HANIN ZAHRI is a senior lecturer in School of Computer and Communication Engineering at Universiti Malaysia Perlis. She received her Ph.D. in Medical Engineering, on Communication and Information System, from University of Yamanashi, Japan in 2013. Her current research interests include natural language processing, machine learning, data mining and data analytics. She teaches programming, software engineering and data analytics course</p>
	<p>R. Prof. Dr. R.Badlishah Ahmad is a Professor in Malaysia. He is Deputy Vice Chancellor (Research and Innovation), Universiti Sultan Zainal Abidin (UniSZA) since 15 March 2017. Graduated PhD (1999) from University of Strathclyde (Scotland, UK). He has supervised more than 40 PhD and MSc students. Specialized and Expertise in Computer and Telecommunication Network Modelling, Embedded System Design and Open Source Software.</p>