Determination of Potential Criminals in Text Analysis: Case of Study

Determinación de Criminales Potenciales en Análisis de Textos: Caso de Estudio

Peter Oropeza[1], José Alberto Hernández Aguilar[1], Alberto Ochoa-Zezzatti[2], Edgar Cossio[3], Julio César Ponce[4]

[1]Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Col. Chamilpa. Cuernavaca, Morelos, C.P. 62209. México.
[2]Universidad Autónoma de Ciudad Juárez
[3]Universidad Enrique Díaz de León
[4]Universidad Autónoma de Aguascalientes

* Correo-e: peteroropeza2@gmail.com, jose_hernandez@uaem.mx, alberto.ochoa@uacj.mx, kofrran@gmail.com, julk_cpg@hotmail.com

KEYWORDS:

ANN, Criminal Activities, Detection & Recognition, Pattern classification, MLP.

ABSTRACT

This research is oriented to classify text using Artificial Neural Networks (ANN) specifically Multilayer Perceptron (MLP) with basic word embedding techniques. The classification consists in determining whether the text has criminal background or not by pattern recognition. The MLP was trained under supervised training and so far with a short range of vocabulary and training records, which each one has a maximum length of 300 words to make the classification process. Analyzing these types of text could help security forces of government, military, etc. to easily detect people who could harm the population and predict possible attacks and prevent them. The developed software needs more word embedding techniques, a larger vocabulary and more training records to be more efficient. The dataset consist of two main classes that are organized as crime and regular type of text.

PALABRAS CLAVE:

RNA, Actividades Criminales, Detección & Reconocimiento, Clasificación de patrones, PMC.

RESUMEN

Esta investigación está orientada a clasificar textos usando Redes Neuronales Artificiales (RNA) específicamente el Perceptron Multicapa (PMC) con Técnicas básicas de palabras embebidas. La clasificación consiste en determinar ya sea que el texto tenga un contexto criminal o no por medio de reconocimiento de patrones. El PMC fue entrenado bajo entrenamiento supervisado y en un rango corto de vocabulario y registros de entrenamiento, cada uno de los cuales tiene una longitud máxima de 300 palabras para hacer procesos de clasificación. Analizar estos tipos de textos podría ayudar a las fuerzas de seguridad del gobierno, a los militares, etc. para fácilmente detectar gente que podría dañar a la población y predecir posibles ataques y prevenirlos. El software desarrollado necesita más técnicas de palabras embebidas, un vocabulario más grande y más registros de entrenamiento para ser más eficiente. El conjunto de datos consiste de dos clases principales que están organizadas como textos de tipo criminal y regular.

Recibido: 3 de mayo 2018• **Aceptado:** 6 de octubre de 2018 • **Publicado en línea:** 28 de febrero de 2019

1. INTRODUCTION

Undoubtedly, humanity has tried to coexist under certain types of regulations to maintain public order, which has always manifested the phenomenon known as crime that is the opposite of regulations and public safety. Considering as much variables as possible to analyze and detect crime threads based on text publications or conversations will be an important tool to predict possible future crimes and prevent them. Crime text analysis relies on a systematic approach for detecting, finding and inclusive predicting crime incidents. The input data consists on text which the MLP classifies in order to certain features or patterns based on an initial vocabulary and the output will be a number 1 in case the input is related to those features related to crime patterns, or a number 0 otherwise. The complex area of criminal data could be an extensive research due to all those hidden or unimaginable related data, which has opened a big door to machine learning to help solving these kinds of problems and find those hidden related patterns, which is useful for all crime investigators personal. The big amount of criminal data hosted in police departments and considering the complexity of relating these data patterns, forces traditional crime analysis methods to become obsolete, due to the amount of human dedicated time to do the research and also it has possible human errors. Considering all those previous variables, is why it’s needed a systematic solution to have a better and efficient approach in crime investigation circumstances.

Artificial Neural Networks techniques can be the key solution. Considering the approach of pattern recognition of an ANN, more researches have joined the subject. The training process is supervised and helps to categorize a context when it’s a crime or not involved.

This research consists of 5 main sections. The first section is the introduction to crime analysis and how ANN has gained territory over this topic. The second section is focused on previous related works to crime analysis. In section three, is explained the methodology used and how it works to classify or learn patterns, as also describes the word embedding technique occupied and the pattern features. Section four presents the results of the MLP. Finally, section 5 presents conclusion and future work.

2. RELATED WORK

Nowadays, a large amount of studies and researches have been accomplished on crime text detection. The result of these investigations is that every day there are new applications to detect and analyze possible crime prospects. In [7] the authors showed a general overview on implementing artificial intelligence crime analysis methods including ANN, Bayesian networks, and genetic algorithms in predicting possible crime events. In [9], the authors worked with ANN for data classification using supervised and unsupervised learning methods. In Mexico there is a great need to detect these types of text due to the high level of violence presented in each state of the republic. Drug traffic has been a main issue in Mexico’s violence, each cartel fight for its own business territory. Often are seen narco-blankets over certain cities, which have threat messages and high levels of violence to overcome. Obviously, these kinds of messages are quickly detected as a violence-crime related message. But the intention is to find possible patterns where is difficult to be detected by human resource and apply efficient models of ANN which can relate these type of data over the internet, conversation, text messages, social media publications, and so on.

3. METHODOLOGY USED

Main architectures of ANN

In general, an Artificial Neural Networks (ANN) can be classified into three parts, named layers, which are known as:

- *Input layer*: This layer is responsible for receiving the inputs to the system.
- *Hidden layer(s)*: This layer is formed by neurons and is responsible for finding all possible patterns for a particular problem and there may be a number of hidden layers, everything will depend on the problem and how much is adjusted for a better approximation.
- *Output layer*: This layer is also formed by neurons and its task is to give a certain resulting value based on the previous layer neuron pattern recognition.

Multiple Layer Feed-Forward Architectures

MLP are formed of one or more hidden neural layers (Fig. 1). They are employed in most cases related to a function approximation, pattern classification, optimization, robotics, and many other useful utilities.

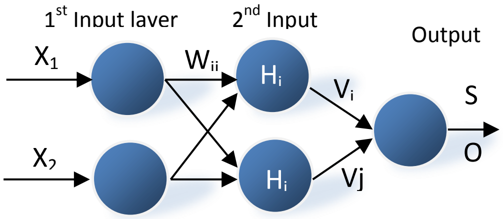


Figure 1. MLP model

Figure 1 shows a MLP feed-forward network, which has an input layer with *n* samples or inputs, it also has two hidden neuron layers and finally, one output neural layer formed with *m* neurons representing the respective output values obtain by the analyzed problem. MLP are a class of ANN who use a feed forward algorithm while training the ANN and during respective process, it will keep updating neurons weights by side with other technique explained later. The number of hidden layers and their amount of neurons depends on a specific problem and how patterns are being recognized.

We will denote h_i as the value the neuron has based on the result of the first step of feed-forward W_{ij} and is represented in equation 1.

$$h_i = f\left(\sum_{i=1}^{n} \sum_{j=1}^{m} x_i * w_{ij}\right) \tag{1}$$

W_{ij} are the weights of the connections between the first input layer and the second one. These weights can be initialized with a random value or with previous data calculated.

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

We can appreciate an f in equation 2 which represent the activation function for the ANN; there is a set of these functions such as sigmoidal, relu and others. Many natural processes and learning curves of complex systems show a temporary progression from a lower level to the beginning, until approaching a climate after a certain time; the transition occurs in a region characterized by a strong intermediate acceleration. The sigmoid function allows describing this evolution. Its graph has a typical "S" shape. The activation function used in this research was the sigmoidal function and can be visualized in equation 2.

V_i is other of the feed-forward algorithm variables. These weights as the same as W_{ij}, they can be initialized with random values or previous data. The output is denoted as O_k and represents the value for the set of entries; the way to calculate O_k is shown in equation 3.

$$O_k = f\left(\sum_{i=1}^{n} h_i * V_i\right) \tag{3}$$

Back-Propagation Algorithm

It should be emphasized that the types of learning are supervised (those that have a set of training inputs and can subsequently perform the classifications) and on the other hand, the unsupervised learning (which does not have a training as input, but rather classifies the data through clustering or other techniques). So is necessary to first train our MLP to later classify data, due to the supervised learning technique used in this research.

Back propagation algorithm starts calculating the error that exist in O_k result of the feed forward algorithm compared to the expected result for those set of entries, denoted by *d*. This error can be calculated as shown in equation 4 and is denoted as *e*.

$$e = d - O_k \tag{4}$$

When calculating *e* if the result is less or minor than a constant defined, it means that is in an acceptable range and the neural net learned this specific pattern; else it would start readjusting all weights backwards in the MLP.

For readjusting V_i weights which are the first weights backwards, is necessary to calculate Δ_k that is a probabilistic value determined as shown in equation 5. Other variable in the scene is α that represents the percentage of learning we want our MLP to converge. Equation 6 explains the way to recalculate V_i weights.

$$\Delta_k = O_k\,(1 - O_k)(d - O_k) \tag{5}$$

$$V_{i+1} = V_i + \propto * h_i * \Delta_k \tag{6}$$

For readjusting W_{ij} weights it's trough equation 8, but first is necessary to calculate Δ_j which is also a probabilistic value calculated as shown in equation 7.

$$\Delta_j = h_j\left(1 - h_j\right)\left(V_j * \Delta_k\right) \qquad (7)$$

$$W_{ij} = W_{(ij-1)} + \propto * h_i * \Delta_j \qquad (8)$$

Once the weight readjusting has finished, we can try and validate if with this new values our MLP can have a better approach to the expected result. This process will continue until it validates the defined error margin.

Input Data Embedding

To feed the MLP is known that is necessary to have numeric values but the main problem is oriented with text analysis, so there are different techniques to give this text a numeric value. This research uses a determined floating number to a specific word as shown in Table 1.

Table 1. Word embedding

Word	Value
Knife	0.1
Robbery	0.2
Drug	0.3
Heroin	0.4
Cocaine	0.5
Gun	0.6
Burn	0.7
Kill	0.8
Kidnap	0.9
Any other word	0

So once having the set of words shown in Table 1, the input data will be provided by a text file and will read the maximum amount of 300 words. After the training session has finished, we will try with a set of text files to classify data and have an output of 1 in case that has criminal intentions or else it will output a 0. The implementation is developed in Matlab.

4. RESULTS

For testing purposes, there were ten experiments which each one has a sentence, and for each sentence there were a result based on the MLP learning session. The testing and result are shown in Table 2, corresponding 1 to those sentences that lead to crime and 0 otherwise.

Table 2. Testing sentences and results

Sentence	Result
A man tried to make a robb	0
A man tried to make a robbery	1
I have a blue dog	0
I want to burn his car	1
My new gun is better than yours	1
Today I bought some cocaine at the store	1
My cousin brought some drugs from the store	1
The senator just got kidnap this morning	1
Carlos wants to cut his eye with the knife	1
Yesterday I lost my purse	0

In Figure 2 is represented the pattern classification or determination when the text involved in the sentence is a crime or not.

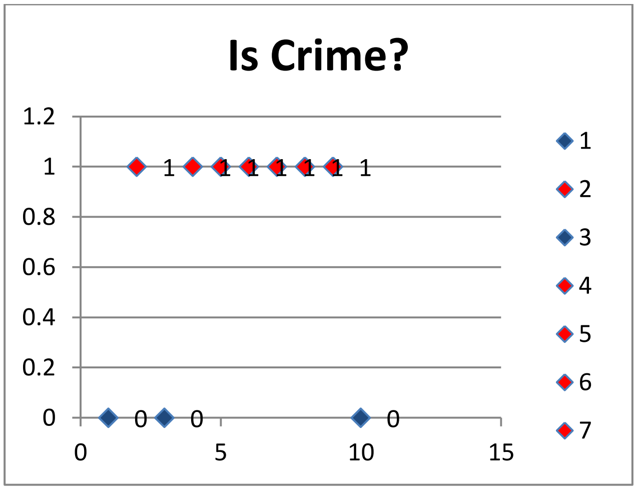


Figure 2. Classification results by the MLP model.

5. CONCLUSIONS AND FUTURE RESEARCH

A Model is developed to detect and predict possible criminals by text analysis using the architecture of Artificial Neural Networks specifically MLP. The results obtained were 95% accurate based on the post-learning session. For future work, we will try on other word embedding techniques, a large dataset of entries for better pattern recognition, as also as possible exit routes in fire evacuation events, based on people's narratives who could reach to their exit routs during the evacuation sinister, focused on the chief's author master's thesis.

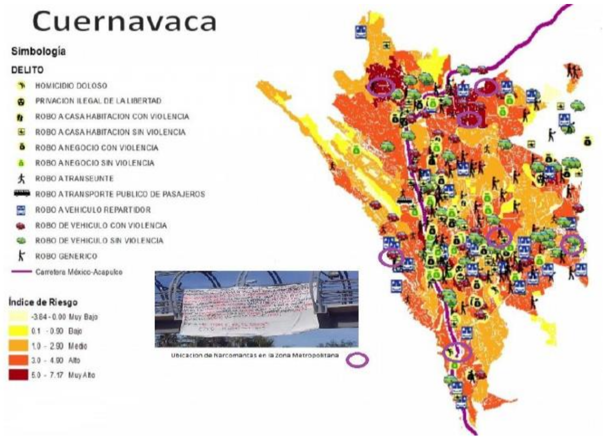

Cuernavaca
Simbología
DELITO
HOMICIDIO DOLOSO
PRIVACION ILEGAL DE LA LIBERTAD
ROBO A CASA HABITACION CON VIOLENCIA
ROBO A CASA HABITACION SIN VIOLENCIA
ROBO A NEGOCIO CON VIOLENCIA
ROBO A NEGOCIO SIN VIOLENCIA
ROBO A TRANSEUNTE
ROBO A TRANSPORTE PUBLICO DE PASAJEROS
ROBO A VEHICULO REPARTIDOR
ROBO DE VEHICULO CON VIOLENCIA
ROBO DE VEHICULO SIN VIOLENCIA
ROBO GENERICO
Carretera México-Acapulco
Índice de Riesgo
-3.84 - 0.00 Muy Bajo
0.1 - 0.90 Bajo
1.0 - 2.90 Medio
3.0 - 4.90 Alto
5.0 - 7.17 Muy Alto
Ubicacion de Narcomantas en la Zona Metropolitana

Figure 3. Risk index and crime preview in Cuernavaca Morelos.

Finally, with the location of each narco message, we propose a geospatial model to represent each scenario and determine future situations related with these kinds of criminal groups. We show in figure 3, this model in a map of Cuernavaca; Morelos in Mexico.

REFERENCES

[1]. Van Banerveld, M., Le-Khac, N.A., Kechadi, MT. (2014) Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation. *In: Dang T.K., Wagner R., Neuhold E., Takizawa M., Küng J., Thoai N. (eds) Future Data and Security Engineering. FDSE 2014.* Lecture Notes in Computer Science, vol 8860. Springer, Cham.

[2]. Zhao, X., Tang, J. Crime in Urban Areas: A Data Mining Perspective. *SIGKDD Explorations*, 2018, 20(1): 1-12.

[3]. Tirthankar Dasgupta, Lipika Dey, Rupsa Saha, Abir Naskar: Automatic Curation and Visualization of Crime Related Information from Incrementally Crawled Multi-source News Reports. COLING (Demos) 2018: 103-107

[4]. Zhao, X., Tang, J.: Modeling Temporal-Spatial Correlations for Crime Prediction. *CIKM 2017*, 2017, 497-506.

[5]. Nakamae, S., Kataoka, S., Tang, C., Pu, Y., Vasilache, S., Saga, S., Shizuki, B. and Takahashi, S. BLE-Based Children's Social Behavior Analysis System for Crime Prevention. *HCI*, 2017, (13), 429-439.

[6]. Zhao, X., Tang, J. Exploring Transfer Learning for Crime Prediction. *ICDM Workshops 2017*, 2017, 1158-1159.

[7]. Oatley, G.C., Zeleznikow, J., Ewart, B.W. *Matching and Predicting Crimes*, In Applications and Innovations in Intelligent Systems XII in Proceedings of AI2004, The Twenty-fourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence. Ann Macintosh, Richard Ellis and Tony Allen Ed. London: Springer, 2004, 19-32.

[8]. Ananyan, S.M. Crime pattern analysis through text mining. *AMCIS 2004*, 2004, 236.

[9]. Adderley, R. W. *The use of data mining techniques in crime trend analysis and offender profiling*, Ph.D. thesis, University of Wolverhampton, Wolverhampton, England, 2007.

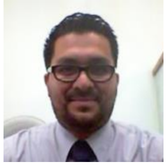
About the authors

Ing. Peter Savier Oropeza Martínez. He is Computer engineer from the Polytechnic University of the State of Morelos. With more than three years of work and scientific experience in private and public institutions in different states of Mexico. He is focused on the area of artificial intelligence, particularly on selected topics such as multiagent simulation, optimization and computational vision. He is winner of first place in the international congress MICAI 2016 at the undergraduate level in the hybrid intelligent systems workshop. Student of the master's degree in optimization and computation applied in the Autonomous University of the State of Morelos.

PhD. Carlos Alberto Ochoa Ortiz. (BSC 1994; Engineering Master, 2000; Ph.D., 2004; Postdoctoral researcher, 2006; Industrial postdoctoral research, 2009). He has written three books and eleven chapters in books related to AI. He has supervised ten Ph.D. theses, 21 Master theses, and 32Bachelor theses. He participated in the organization of conferences such as HAIS’07, HAIS’08, ENC’06, ENC’07, ENC’08, MICAI’09, MICAI’10 and MICAI’11. His research interests include evolutionary computation, natural processing language, and social data mining. He is member of the Mexican National Researchers System Level 2.

PhD. José Alberto Hernández Aguilar. In 2008 he obtained the degree of Doctor of Engineering and Applied Sciences from the Research Center in Engineering and Applied Sciences of the UAEM. His professional experience has been oriented to the development of information systems oriented to decision-making, information analysis through data mining and has recently ventured into the implementation of optimization algorithms in GPU's.

PhD. Edgar Gonzalo Cossio Franco. He received his PhD in computer systems in the Universidad Da Vinci (UDV). Master in Software Engineering from the Universidad del Valle de Atemajac (UNIVA) Guadalajara in 2011. His research interest is the parallel computing, software engineering, bio-inspired algorithms. He is part time professor in the Enrique Díaz de León University.



PhD. Julio Cesar Ponce Gallegos. Received the B.S. degree in computer system engineering from the Universidad Autónoma de Aguascalientes in 2003, the M.S. degree in computer sciences from the Universidad Autónoma de Aguascalientes in 2007, and the PhD. Degree in computer sciences from the Universidad Autónoma de Aguascalientes in 2010. He is currently a professor in the Universidad Autónoma de Aguascalientes. His research interests include Evolutionary Computation, Data Mining, Software Engineering and Learning Objects.