Classification of Digitized Documents Applying Neural Networks

Amelec Viloria, Noel Varela, Omar Bonerge Pineda Lezama, Nataly Orellano Llinás, Yasmin Flores, Hugo Hernández Palma, Carlos Vargas Mercado and Freddy Marín-González

Abstract The exponential increase of the information available in digital format during the last years and the expectations of future growth make it necessary for the organization of information in order to improve the search and access to relevant data. For this reason, it is important to research and implement an automatic text classification system that allows the organization of documents according to their corresponding category by using neural networks with supervised learning. In such a way, a faster process can be carried out in a timely and cost-efficient way. The criteria for classifying documents are based on the defined categories.

Keywords Text categorization · Artificial neural networks · Multilayer perceptron

A. Viloria (✉) · N. Varela · F. Marín-González
Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
e-mail: aviloria7@cuc.edu.co

N. Varela
e-mail: nvarela2@cuc.edu.co

F. Marín-González
e-mail: fmarin1@cuc.edu.co

O. B. P. Lezama
Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

N. O. Llinás · Y. Flores
Corporación Universitaria Minuto de Dios—UNIMINUTO, Barranquilla, Colombia
e-mail: nataly.orellano@uniminuto.edu

Y. Flores
e-mail: yasmin.flores@uniminuto.edu

H. H. Palma · C. V. Mercado
Corporación Universitaria Latinoamericana, Barranquilla, Colombia
e-mail: hhernandez@ul.edu.co

C. V. Mercado
e-mail: cvargas@ul.edu.co

V. Bindhu et al. (eds.), *International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering 637,
https://doi.org/10.1007/978-981-15-2612-1_20

1 Introduction

The automatic classification of documents has gained increasing research interest in recent times, since the exponential increase of available information in digital format over the last few years, and expectations of future growth make it necessary to organize all this content in order to improve the search and access to information, which has become a difficult task by means of the manual classification [1]. To this end, the development of an intelligent system for the automatic classification of texts is necessary.

Since the development of science and technology presents an accelerated advance, the information in each knowledge area increases exponentially, and its treatment and storage become more complex. The explosive growth of information available in digital documents in the area of information technology and systems has made it necessary to develop new tools and instruments that facilitate the conduct of search processes in an efficient and effective way, as well as the management of these resources. In order to facilitate the search for information, documents are often categorized into a limited set of classes or categories. These classes represent specific areas of knowledge and are generally consolidated by experts [2, 3].

The context of this study aims to create an intelligent system that allows documents to be automatically categorized using expert systems [4] in such a way that a faster process is carried out with less time and cost. For the development of the research, data available from the University of Mumbai repository was used in different formats (Microsoft Word, PDF, plain text) that have been used to train and subsequently evaluate the results obtained in each training group. It is important to indicate that the classification of the documents will be categorized into 14 groups which are physics, mathematics, social sciences, natural sciences, art, economy, education, engineering, environment, medicine, juridical, psychology, language and diverse.

2 Method

2.1 Sample Data

The data was collected from the repository of Scientific Journals of the University of Mumbai in India [5]. This repository has a large number of digitized documents such as 1,254,325 documents composed of: full text scientific articles, scientific journals and fascicles.

As stated above, one of the reasons for using this repository is that its metadata has a complete structure such as author, title, keywords, publication, URL, among others. Each of these data provides more information about the document that is why when selecting the repository, the corpus metadata was considered since it will be of great help for classification purposes [6].

Table 1 Metadata used for a document

Authors
Title
Description
Date of publication
Keywords
Language
Url

Metadata is structured and coded data describing characteristics of instances containing information to help identify, discover, evaluate and manage. In other words, metadata is data about the data, and they will be extracted and used in the identification of documents after the analysis of the repositories. Table 1 shows these data.

2.2 Network Architecture Design

Since the study requires the construction of a previously supervised multilayer network, the network outputs must be previously known. This network is applied to the classification of documents into 14 categories which are: physics, mathematics, social sciences, natural sciences, art, economics, education, engineering, environment, medicine, law, psychology, language and diverse. These categories are mentioned after an exhaustive study and review of each of the documents in the repository [7].

In order to determine the belonging of a document to a certain category, a vocabulary had to be assigned to each of them, that is, each category has subcategories that best describe the category. For this purpose, the process includes the Dewey Classification System [8], which constitutes structured lists of terms (concepts) that represent, in a univocal way, the conceptual content of the documents and is a system that quantifies the relevance of a term to describe a category.

To conclude the category, the category–glossary relationship is used, i.e., the network entry is compared with each glossary or vocabulary of each category to assign weights to each category to determine the category to which it corresponds; see Fig. 1.

Figure 2 shows the structure of the network and its connections, the input layers will be represented by title data, keywords and description, and the output layer will be made up of the 14 categories. The final model corresponds to a network with 14 inputs and 14 outputs.

```
public String[] neuralNetworkInput(List<String> words){
    String[] input = {"0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"};

    for (String word : words) {
      Glossary gosa = Glossary List Get Glossary by Word (word);
      if (gosa!=null) {
        List<Categories> cats = gosa.getCategories();
      for (Categories categories : cats) {
        int index = Integer.parseInt(""+Categories.getId())-1;
        float aux = Float.parseFloat(input[index]);
        if (aux!=0) {
          aux = (float) (aux + 0.1 - 0.01);
        }else{
          aux = (float) (aux + 0.1);
        }
        input[index]= ""+aux;
      }
      }
    }
    return input;
  }
```

Fig. 1 Algorithm for weight allocation

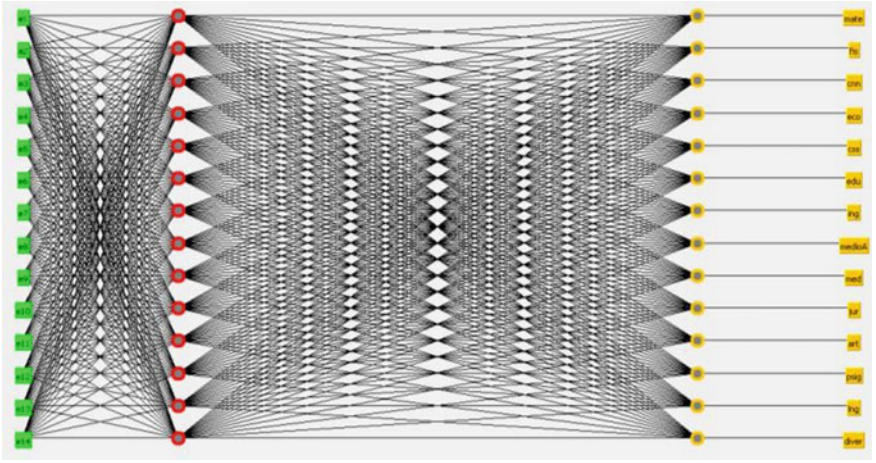

Fig. 2 Result of the network in Weka

2.3 Learning from the Network

Supervised learning is characterized by knowing how the network should respond to a given input. In this way, the desired output is compared with the mains output, and if there are discrepancies, the weights are iteratively adjusted. Thus, the learning stage aims to minimize the error between the output provided by the network and the desired or true output [9, 10].

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1152               97.4619 %
Incorrectly Classified Instances        30                2.5381 %
Kappa statistic                          0.9717
Mean absolute error                      0.011
Root mean squared error                  0.0629
Relative absolute error                  8.5963 %
Root relative squared error             24.8401 %
```

Fig. 3 Sample 1 result

3 Results

These results are based on the two training sets of 2000 and 4000 documents. Then, the results of each are presented and discussed. Weka (Waikato Environment for Knowledge Analysis) [11, 12] was used to develop the corresponding training tests to classify the texts according to the categories, which allows the verification and constancy of the results of the correct and bad classification that may be generated.

3.1 Sample 1

Figure 3 shows the first set of data that consists of 2000 documents, which results present a high percentage of correct classification, indicating that there is a good classification. Then, their results are shown in percentages of classification versatility and margin of error.

When analyzing the sample 1, the instances of correct classification show 97% of well-classified documents and a minimum error margin of 3%, indicating a greater range of correct classification of documents.

In Fig. 4, the confusion matrix shows the type of correct and incorrect predictions about the set of documents. It makes it possible to understand how the network makes a mistake when trying to classify the new set of documents. In the graph of this matrix, the correct predictions are represented on the diagonal [13].

3.2 Sample 2

Figure 5 shows that the second set of data included 4.000 documents. As the number of documents increased, it was necessary to add new words to the vocabulary in order to keep the assertiveness margin from decaying.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   <-- classified as
 164   0   0   0   0   0   0   0   0   0   0   0   0   0 |   a = mate
   0 198   0   0   0   0   0   0   0   0   0   0   0   0 |   b = fis
   0   0  98   0   0   0   0   0   0   0   0   0   0   0 |   c = cnn
   0   0   0 120   0   0   0   0   0   0   0   0   0   1 |   d = eco
   0   0   0   1 141   0   0   0   0   0   0   0   0   0 |   e = css
   0   1   0   0   0 109   0   0   0   0   0   0   0   0 |   f = edu
   0   2   1   0   0   0  13   0   0   0   0   1   0   0 |   g = ing
   1   0   1   0   0   0   0  69   0   0   0   0   1   1 |   h = medioA
   0   1   1   0   0   0   0   0  24   0   2   0   0   0 |   i = med
   0   0   0   0   1   1   0   0   0  51   0   0   0   2 |   j = jur
   0   0   1   0   2   0   0   0   0   0  20   0   0   0 |   k = art
   2   2   2   0   0   0   0   0   1   0   0  62   0   0 |   l = psig
   0   0   0   0   0   0   0   0   0   0   0   0  21   0 |   m = lng
   1   0   0   0   0   0   0   0   0   0   0   0   0  62 |   n = diver
```

Fig. 4 Confusion matrix of sample 1

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        2223               89.3489 %
Incorrectly Classified Instances       265               10.6511 %
Kappa statistic                          0.8819
Mean absolute error                      0.023
Root mean squared error                  0.1109
Relative absolute error                 17.8743 %
Root relative squared error             43.7063 %
```

Fig. 5 Results of sample 2

The result of this new training group shows a value of 90% of correct instances and with a margin of error of 10%. Unlike the first group, the margin of assertiveness was reduced by a minimum percentage due to the large number of documents to be analyzed, that is, as the number of documents to be analyzed increases, the vocabulary should be increased.

As shown in Fig. 6, the confusion matrix identifies the type of correct and incorrect predictions about the data set. It makes it possible to understand in what sense the network is mistaken when trying to classify the new texts [14].

```
== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   <-- classified as
 153   0   4   1   2   4   0   0   0   0   1   0   0   0 |   a = mate
   1 198   2   4   0   3   6   1   0   1   5   0   2   0 |   b = fis
   2   0 121   3   2   5   1   8   0   0   2   1   1   0 |   c = cnn
   1   0   1 113   2   4   0   0   1   2   2   1   0   0 |   d = eco
   3   0   0   2 146   8   1   2   0   3   7   1   1   0 |   e = css
   0   0   1   1   1 357   2   1   1   3   4   0   2   0 |   f = edu
   4   4   1   1   2   9 354   3   1   2   7   1   5   0 |   g = ing
   2   2   3   2   3   4   4 267   3   0   2   0   6   1 |   h = medioA
   6   0   3   1   0   3   2   2 165   0   1   0   0   0 |   i = med
   4   0   0   2   2   4   1   0   0 104   4   1   1   0 |   j = jur
   3   1   0   0   0   2   1   2   0   0 101   0   2   0 |   k = art
   5   1   3   0   0   2   1   1   0   2   1  22   1   0 |   l = psig
   0   0   1   1   0   1   0   0   0   0   0   0 120   0 |   m = lng
   1   0   1   1   0   1   0   0   0   0   1   0   0   2 |   n = diver
```

Fig. 6 Confusion matrix of sample 2

4 Conclusions

The use of the Weka tool allows the execution of training tests of neural networks with the purpose of predicting the area of belonging of a text [15]. To ensure a good classification of documents, it was necessary to increase a greater number of words to the vocabulary in sample 2 in order to give a better accuracy in the classification of documents.

Correctly defining the vocabulary of each of the categories makes the classification to have a good percentage of successes, correctly assigning the document to its category, in such a way that the performance of the network accuracy improves according to the size of the vocabulary. The use of metadata helped achieving better results in the representation, localization and retrieval of electronic resources.

References

1. Vásquez, A.C., Lazo, O.R., Agnelli, R.C.: Categorización de Textos mediante Máquinas de Soporte Vectorial. Revistas Signos, pp. 1–24 (2011)
2. Mendoza, M., Ortiz, I., Rojas, V.: Categorización de texto en bases documentales a partir de modelos computacionales liviano. Revista de investigación de Sistemas e Informática. **10**(1), 2–12 (2013)
3. Pérez, P.M., Colarte, J. (Feb, 2007) Multimedia para discapacitados. Presentada en: Congreso y Feria Internacional Informática 2007 (en línea). Disponible en: http://www.informaticabana.cu/eventovirtual/educacion/discapacitados.pdf
4. Bechara, J.E.A., Cruz, J.C.T., Ceballos, H.V.: Predicciones de modelos econométricos y redes neuronales: el caso de la acción de SURAMINV. Semestre Económico Universidad de Medellín. **12**(25), 95–109 (2009). Available from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120634620090003000007&lng=en&nrm=i-so. Access on 07 Aug 2017

5. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487–499 (1994)
6. Hahsler, M., Karpienko, R.: Visualizing association rules in hierarchical groups. J. Bus. Econ. **87**, 317–335 (2017)
7. Silverstein, C., Brin, S., Motwani, R., Ullman, J.: Scalable techniques for mining causal structures. Data Min. Knowl. Discov. **4**(2–3), 163–192 (2000)
8. Amelec, Viloria, Carmen, Vasquez: Relationship between variables of performance social and financial of microfinance institutions. Adv. Sci. Lett. **21**(6), 1931–1934 (2015)
9. Viloriaa, A., Lezamab, O.B.P.: Improvements for determining the number of clusters in k-means for innovation databases in SMEs. Procedia Comput. Sci. **151**, 1201–1206 (2019)
10. Kamatkar, S.J., Kamble, A., Viloria, A., Hernández-Fernandez, L., Cali, E. G.: Database performance tuning and query optimization. In: International Conference on Data Mining and Big Data. Springer, Cham, pp. 3–11 (2018)
11. Viloria, A., et al.: Integration of data mining techniques to postgreSQL database manager system. Procedia Comput. Sci. **155**, 575–580 (2019)
12. Lanzarini, L., Villa Monte, A., Aquino, G., De Giusti, A.: Obtaining classification rules using lvqPSO In: Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science. vol. 6433, pp. 183–193. Springer, Berlin, Heidelberg (2015)
13. Borja-Borja, M.G.: Algoritmo de Entrenamiento de una Neurona Artificial Perceptrón para Reconocimiento de Caracteres, Aplicando Principios Heurísticos. Revista ECIPerú. **6**(1), 4 (2009)
14. Alonso, M.Á.L.: Las estructuras conceptuales de representación del conocimiento en internet. Scire: representación y organización del conocimiento. **6**(1), 107–123 (2000)
15. Torrez Torrez, E.D.: Sistema inteligente para la detección de conversaciones con posible contenido pedofílico basado en redes neuronales, Doctoral dissertation (2018)