Promoting Social Media Dissemination of Digital Images Through CBR-Based Tag Recommendation

Lucía Martín-Gómez*, Javier Pérez-Marcos, Rebeca Cordero-Gutiérrez, Daniel H. de la Iglesia

Department of Computer Science, Pontifical Univeristy of Salamanca C/Compañía, 5, 37002
Salamanca (Spain)

Received 15 April 2022 | Accepted 29 June 2022 | Early Access 16 September 2022

Abstract

Multimedia content has become an essential tool to share knowledge, sell products or disseminate messages. Some social networks use multimedia content to promote information and create social communities. In order to increase the impact of the digital content, those images or videos are labeled with different words, denominated tags. In this paper, we propose a recommender system which analyzes multimedia content and suggests tags to maximize its influence in the social community. It implements a Case-Based Reasoning architecture (CBR), which allows to learn from previous tagged content. The system has been evaluated through cross fold validation with a training and validation sets carefully constructed and extracted from Instagram. The results demonstrate that the system can suggest good options to label our image and maximize the influence of the multimedia content.

Keywords

Artificial Intelligence, Digital Image Processing, Recommender System, Social Media, Tagging.

DOI: 10.9781/ijimai.2022.09.002

I. Introduction

The information is the essence of any communication support. The data offered to the individuals provokes a reaction to consume them, either because of its quality, its originality or the way in which it is told. In the particular case of Internet, this issue becomes fundamental for the knowledge sharing. With regard to a digital entity (i.e. a webpage, a social profile or a digital product) it is difficult to achieve the objectives for which it is created when interesting information for its visitors is not present or it is poorly ranked.

As we live in a multimedia world, the information has stopped being limited to a text, but to a mixture of digital objects that allow us to transmit a message. Therefore, text is supported by other visual elements, denominated multimedia content, that serve to draw the attention of our audience. Thus, multimedia content becomes a more attractive alternative for those users who prefer this type of support instead of reading on the website. In this way, the information is mainly based on images and conveys the message we are trying to promote. Furthermore, the continuous increase of users connected to the Web requires new methods to maximize the impact of information dissemination.

In particular, social networks have taken advantage of the power of multimedia content to promote their data. Moreover, some social networks such as Instagram or TikTok have made the multimedia content their essence to survive and expand in the Internet.

Instagram[1] is a social network to upload photos and videos. The users can also apply photographic effects such as filters or frames, add text, gifs and stickers to their posts or create compilations of several short video fragments. Despite its recent birth in 2010, its concept has been rapidly accepted by the society, and by the end of 2021 it had more than 2 billion active users [1].

Images in many social networks are promoted through the use of tags, which consist of short words that somehow describe the content or the purpose of the picture. Tags are essential for the correct dissemination of multimedia content through the social platform. The tagging of multimedia content to categorize publications by subject matter on the social media is done through the so-called hashtags, so henceforth tags or hashtags will be referred to indistinctly. There are a series of metrics that are calculated based on the interactions of other users with the post (like, share the post, write a comment, save the post...) Thus, the selection of the words to tag the multimedia content becomes essential to augment the visibility of the image and therefore the user. However, despite of different proposals to tag different kind of content [2]–[4] there is no standard method in social networks to know beforehand which words are better to optimize the impact of the image.

In this paper, we present a recommender system that suggests tags to promote a digital image submitted to social networks. In order to improve the performance, the recommender consists of a Case-Based Reasoning (CBR) architecture, which is able to learn from previous experiences to obtain better results in the future. Initially, the memory of the system is previously populated with image features obtained from a set of photos uploaded to Instagram and their associated tags. Then, the system can recommend tags for a particular image manually selected.

For this purpose, the main features of the image are extracted and analyzed. With these features, we obtain a map which is compared with previous images stored in the memory and selected those ones which are applicable due to their similarity. Finally, a set of words are

[1] https://www.instagram.com/

* Corresponding author.

E-mail address: lmartingo@upsa.es

chosen following a set of rules created. Since this process coincides with the theoretical approach of a CBR and taking into account that the literature shows that this type of systems obtain very good results in tagging and recommendation problems, our proposal implements a CBR for the image tagging task.

The experiments aim to demonstrate the performance of the system independently of the dataset, by applying a cross validation. Additionally, we aim to prove that CBR can suggest good tags to label multimedia content. For this purpose, a comparison between the CBR system and other regular classification systems was performed.

The remainder of the paper is structured as follows. Section II briefly describes recent works related to the recommendation in multimedia content. Section III discusses the techniques used for image processing. Section IV provides the technical details of the recommender system. Sections V and VI detail the case study in Instagram and experiments carried out to validate this proposal and discusses the preliminary results. Finally, Section VII exposes the main contributions of this work.

II. Recommender Systems for Multimedia Content

Feature extraction of multimedia content has been deeply explored to create recommender systems. In music, [5] applies a set of boosted classifiers to map audio features onto social tags collected from the Web. The resulting automatic tags are part of a social recommender. [6] predicts potentially interesting and unknown music based on an analysis of musical features of musical tracks. [7] proposes a recommender system to suggest music by applying data mining techniques with information about its content and the context. [8] proposes a model for recommendation to predict the latent factors from music audio when they cannot be obtained from usage data. [9] learns features from audio content and makes personalized recommendations. In images, [10] presents an analyzer to extract features from images for recommendation purposes. [11] proposes a progressive image search and recommendation system, which incorporates the auto-interpretation and user behavior.

There are some approaches that suggest new tags for specific digital content. For instance, [2] present TagAssist, a recommender system that recommend tags for posts by applying a Case Based Reasoning (CBR) architecture. [12] propose a strategy that enables a content-based recommender to infer user interests by applying machine learning techniques both on the “official” item descriptions provided by a publisher, and on tags which users adopt to annotate relevant items. More recently, [3] proposes a new framework which makes recommendation of tag-based multimedia recipe. [4] framework that is able to utilize knowledge over the Linked Open Data (LOD) cloud to recommend context-based services to users. However, these proposals do not consider the multimedia content for the calculation of new tags.

Additionally, recommender systems commonly makes use of learning techniques, specially CBR, when multimedia content is involved. [13] makes use of a CBR to tag emotions from facial expressions. [14] presents a new recommender with a CBR that exploits audio and tagging knowledge using a hybrid representation and adding semantic knowledge extracted from the tags of similar music tracks. [15] presents a medical CBR system with a knowledge-based recommendation, which analyzes image and text from patient health records. [16] presents a CBR that exploits nuclear image features to retrieve the cases that are the most similar to the new image test and to compute the most probable diagnoses. [17] develops a CBR System for face recognition under partial occlusion. All the proposals obtained very successful results, even when compared with other similar techniques used for the same purposes.

In terms of automatic labeling of multimedia information, the results obtained are not fully satisfactory, so most proposals opt for human-machine collaboration for more accurate and efficient multimedia tagging [18]. [19] proposes a D2-clustering-based method that represents the multimedia content by bags of weighted vectors. On the other hand, [20] describes a scalable algorithm that considers the use of matching labels in images with similar characteristics by accumulating votes from visually similar neighbors. Other proposals, such as [21], combine historical image tagging information and metadata into an adaptive factorization model that applies transfer learning and deep learning image classification techniques.

Another important concept in tagging and recommendation systems for multimedia content is that of folksonomies, which is a system for assigning tags to elements by users [22]. Thus, the assignment of a tag for a particular content is influenced by the general criteria of the users [23]. Some authors have already made use of folksonomies in multimedia content tagging problems. When social tagging generates folksonomies in image-related content, these are called visual folksonomies. For instance, [24] describes some techniques for automatic image tagging that take benefit of collaboratively tagged image databases and [25] formulates image tag recommendation as a maximum a posteriori (MAP) problem, making use of a visual folksonomy.

As it is demonstrated with some recent related work, CBR can get very good results in tagging and recommender systems [26], [27]. This careful literature review leads us to build a new CBR-based assistive tagging system capable of analyzing multimedia content and suggesting tags to maximize its impact on social networks.

III. Image Feature Extraction

As we stated in the introduction, the present work aims to recommend tags in order to promote multimedia content in a social network. Recommenders usually retrieve information from different data provided by the user, such as personal information or navigation data. Companies Amazon, Spotify or TripAdvisor find that information as essential for the correct operation of their recommenders [28].

It becomes clear that the feature extraction and its analysis is needed for the recommendation process. Therefore, starting point of this system is the extraction of the main features of a digital image. At this step, if the multimedia content submitted to the social network is a video file, its representative image (video thumbnail) is considered. Then, the extracted visual descriptors are compared with those of other images’ to search for similar experiences.

An image is described by the shape and layout of its elements, as well as its colors. These types of descriptors are the main features considered in image preprocessing tasks. Due to their nature, these parameters are usually divided into two categories: shape and disposition, and color. This taxonomy permits to address different problems in which not every feature of the image should be involved. As an example we can cite the identification of a particular object in an image, which does not consider the particular colors, or the search of color histogram, which does not depend on the shape and disposition of figures in a picture.

In this sense, there are different techniques to detect the shape and disposition of the elements that are part of an image. For this particular problem, we focus the algorithms that capture invariant image descriptors, as we aim to detect similarities even if the image is rotated or transformed. Among the different proposals, Scale-Invariant Feature Transform (SIFT) [29] is of particular interest due to their successful results in different problems. Some versions of SIFT, such as Speeded Up Robust Features (SURF) [30] and Oriented FAST and

Rotated BRIEF (ORB) [31], were proposed to improve the computation time of the original algorithm, which in some cases, can become a little high. In order to also reduce the dimensionality of the problem and improve the times by grouping the descriptors into clusters, Bag of Visual Words (BoVW) [32] is frequently applied.

To analyze the color of a digital image, color histograms have been long-established in this field [33]. This technique obtains the color distribution based on the values of the pixels. However, the immense amount of pixels, and therefore, colors that compose a digital image involves a great difficulty in the color extraction process. As a solution, color quantization procedure is widely used to reduce the number of colors in an image, extracting only the most representative ones [34]. The color quantization can be implemented following different algorithms, such as clustering, k-means or neural networks. It is essential to analyze each proposal and select which one can better adapt to solve our problem.

Moreover, transfer learning involves the exploitation of learning outcomes to a related task [35]. In neural networks, transfer learning is applied by obtaining the weights of the layers prior to the classification of a model already trained for a similar problem, which are given as feature vectors called embeddings. [36] applies this technique with the Inception-v3 model [37] to classify different Instagram influencer profiles according to their interests or posting topics (i.e. fashion or beauty). Thus, this technique reduces the computational cost of feature extraction, optimizes the results by working with already validated data related to the problem to be treated and makes the feature extraction phase much more efficient [38].

After a careful examination of the literature and taking into account the main objective of this work in which the execution time is not critical, an original version of the SIFT has been considered. In order to reduce the dimensionality of the problem, BoVW is also applied in the image feature extraction process. For the color extraction, a color histogram was obtained to represent chromatic information from the images. Additionally, in this proposal, transfer learning will also be applied based on the embeddings obtained from the Inception-v3 model [37].

IV. Recommender System

This paper proposes a CBR-based recommender system that relies on image metadata to propose tags for disseminating multimedia content. Thus, the image is the input to a system that generates, as a recommendation, a set of tags appropriate to its content.

Our proposal is based on the CBR methodology combined with state-of-the-art techniques within the field of deep learning. Unlike other content-based recommender systems, this work uses Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) in two different ways: on the one hand they are used to infer a Case-Representation feature space, and on the other hand to define a hashtag latent space embedding. The definition of a Case-Representation feature space allows the CBR stages of retrieval and reuse to use a distance-based recommendation. The creation of a latent space of hashtags allows that given a hashtag to recommend we can obtain those closest ones that are more likely to be used in combination.

The Case-Base of our CBR is formed by Case-Representations obtained from an initial set of images labeled with hashtags. This set of images will be used to train the Case Representation Neural Network and subsequently to initialize the Case-Base, as shown in Fig. 1. The idea of using a Case Representation Neural Network (CRNN) is to obtain a Case-Representation space where the items on which the same tags are used are close and at the same time far away from those that do not use them. In this way, we will be able to use distances in the retrieval phase, such as the squared distance, on the cases to obtain the closest ones [17]. The CRNN is formed by the embedding layer of a Siamese Neural Network using Triplet Loss [39].

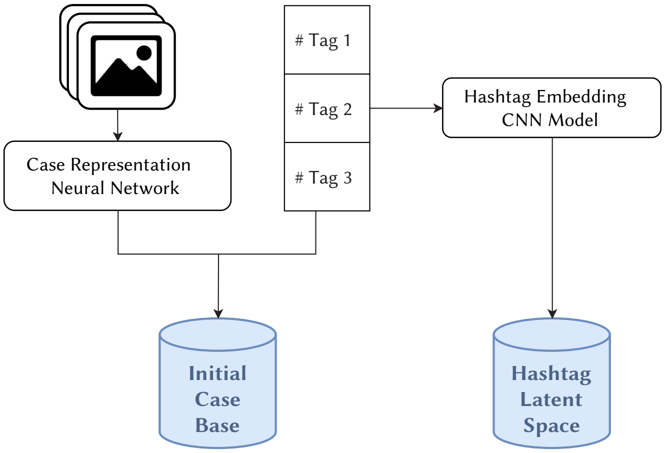
# Tag 1
# Tag 2
# Tag 3
Hashtag Embedding CNN Model
Case Representation Neural Network
Initial Case Base
Hashtag Latent Space

Fig. 1. Overview of the case-base inizialization and hashtag latent space creation process.

The hashtags latent space is formed by the Neural Network Embeddings of the hashtags. This initial latent space is obtained from the same CRNN training set as shown in Fig. 1, but in this case using the hashtags. When recommending hashtags, the reuse phase of the CBR will propose the hashtag that best suits the new case. However, more than one hashtag per image must be recommended. To this end, once the proposed solution has been obtained, the latent space of hashtags will be searched for those that come closest to the proposal. To obtain this latent space of hashtags, a DNN has been trained to obtain the embeddings, so that those hashtags that are often used together are close in this latent space, while those that are never used together are far away.

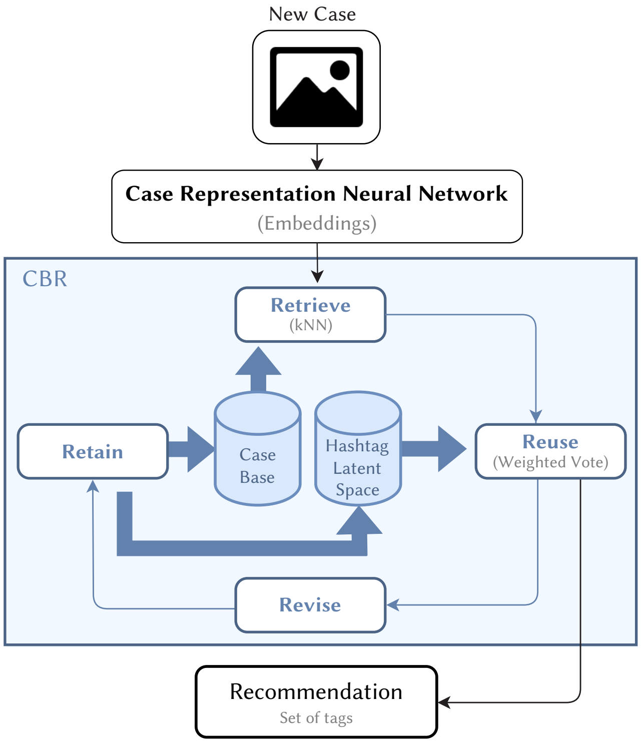

Fig. 2. Overview of the proposed CBR-base recommender system.

In order to be able to make recommendations, the Case-Base must have been initialized and the latent space of hashtags must have been defined. To make a recommendation on a new image, its Case-Representation must first be obtained. To do this, the image is passed through the CRNN to obtain its embedding. In the retrieve phase, the most relevant cases of the Case-Base are obtained from the Case-Representation of the new case and by means of the squared distance. In the following reuse phase, the proposed hashtag is obtained by a weighted vote of the retrieved cases. From the obtained proposal, the k closest hashtags are searched in the latent space of hashtags, forming together the recommendation. In the revision phase, we check which of the hashtags finally used were not recommended, to store the Case-Representation together with the non-recommended hashtags in the retention phase. These last two phases allow our system to be able to adapt to new hashtags and learn from users' tagging habits. The complete cycle of the proposed CBR can be seen in Fig. 2.

A. **Retrieve**: Getting the Best Tags

In the retrieval phase of a CBR, the system recovers from the Case-Base the cases most similar to the Case-Representation of the new case. The Case-Representation of our proposed system for an image x is an embedding $f(x)$ such that in a feature space R^d the squared distance of identically labeled images is small, while for differently labeled images it is large.

1. Image Embedding Network Architecture

The CNN architecture responsible for the image embedding can be seen in Fig. 3. The first stages of the network are reused from another pre-trained network for the task of classifying images. This technique is known as transfer learning. In this way we can obtain a feature vector from an image, without the need to retrain it. The pre-trained network for our proposal is Inception-v3 due to its outstanding performance [37]. In order to use a pre-trained network, the final stages in charge of classification must be removed. In the case of Inception-v3 we have removed the last two layers (an Average-Pooling pre-classification layer and a fully-connected layer), leaving a final layer of (8 x 8 x 2048) components.

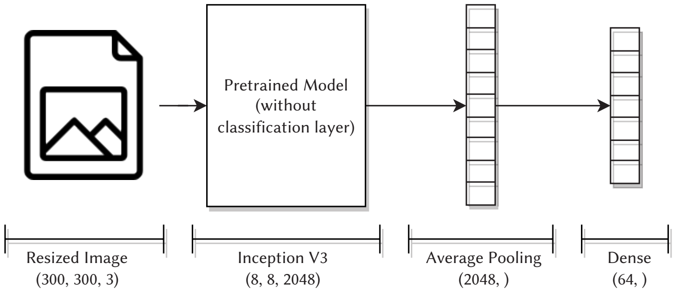

Fig. 3. Overview of the Case Representation Neural Network architecture.

The next two layers added to the modified pre-trained network are in charge of obtaining the embeddings of the images. The first one consists of a 1-dimensional Average-Pooling layer. This layer allows to reduce the dimensionality of the feature maps of the previous layer, making it more robust to changes in the positions of the image features. The second layer consists of a Dense layer, using the hyperbolic tangent as an activation function. The output of this layer will be the embedding of the image, so the size of the layer will determine the size of the embedding. In our case we have chosen a size of 64 components for this layer.

2. Siamese Network With Triplet Loss Architecture

As mentioned above, a Case-Representation has to be a set of features from which we can at the retrieval stage recover those cases of the Case-Base that are most similar. It is important that the set of features forming the Case-Representation represents an image in an embedded space in such a way that semantically related images are metrically close. For this purpose, a Siamese Neural Network has been used together with the Triplet Loss as a cost function.

The Siamese Neural Network structure consists of two branches formed by the same neural network model that share the weights and parameters [40]. During training, the network is fed with image pairs. The objective of this network is to learn the optimal features of the images in such a way that related images are pulled closer while those that are not pushed away. To optimize the neural network, a cost function capable of making a distinguish between pair is used defined in (1).

$$\mathcal{L} = \frac{1}{2} l D^2 + \frac{1}{2}(1-l)\{\mathrm{m}\ \ (0, m-D)\}^2 \tag{1}$$

Where l is a binary label selecting whether the input pair consisting of image x_1 and x_2 is a positive ($l = 1$) or negative ($l = 0$), m>0 is the margin for dissimilar pairs and $D = \|f(x_1) - f(x_2)\|_2$ is the Euclidean distance between feature vectors $f(x_1)$ and $f(x_2)$ of input images x_1 and x_2.

The neural network used in the proposed system is a variation of the Siamese Neural Network called Triplet Neural Network [39]. Unlike the Siamese Neural Network, this one consists of three branches with the same neural network model, sharing the same weights and features. The input of this network is formed by a triplet of objects. While in a Siamese Neural Network the pairs of objects could be related or unrelated, in a Triplet Neural Network one of the objects is the anchor, while of the remaining two one is related to the anchor (positive) and the other is unrelated (negative). Formally, for the triplet (x^a, x^p, x^n) one (x^a is the anchor, x^p is the positive and x^a is the negative) has that r(x^a, x^p)>r (x^a, x^n) where $r(.)$ is a similarity measure. The cost function of the Triplet Neural Network is the Triplet Loss Function. We want the image x_i^a to be closer to all images x_i^p than to any of the images x_i^n, as shown in (2).

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \\ \forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \mathcal{T} \tag{2}$$

Where $f(x_i)$ is the embedding of an image x_i, α is a margin that is enforced between positive and negative pairs, and T is the set of all possible triplets. Then, the network cost function to be minimized is described in (3).

$$\mathcal{L} = \sum_i^n \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \tag{3}$$

The overall architecture of the Triplet Neural Network is shown in Fig. 4.

B. **Reuse**: Suggesting Tags Based on the Experience

In the reuse phase, the best solutions are suggested from the cases retrieved in the previous stage. The most common method for this purpose is the weighted vote of the solutions proposed by the cases using their distance to the new case. In the case of a multi-label system such as the proposed one, one can either retrieve the k most voted solutions or use multi-label implementations of the Nearest Neighbors algorithm. Our proposal is to use the most voted solution, and then to search in the latent space for their k closest solutions using the Euclidean distance.

Given that the solution space is large, and that the problem to be solved such as recommending hashtags for a folksonomy is complex due to problems such as the user's freedom in defining the hashtags and the constant evolution of these hashtags, an alternative is

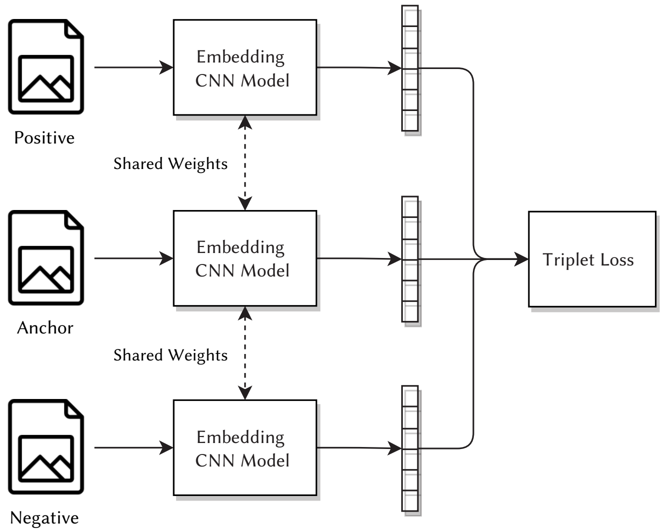

Fig. 4. Overview of the Siamese Neural Network architecture using triplet loss function.

proposed. Our approach consists of reducing the solution space to a set of semantically grouped clusters. In this way, hashtags belonging to the clusters obtained in the CBR reuse phase will be recommended.

1. Hashtags Latent Space

The hashtags recommended by our system are retrieved from a latent label space from the proposed solution. This latent space is formed by the embeddings of the hashtags built in the initialization of the system from the Case-Base. The label embeddings are obtained from a DNN. This DNN takes as input a tag and the ID of an image and identifies whether the tag and the ID are related or not. That is, given a tag t_i and an image i_i, $r(t_i, i_i) = 1$ if they are related, and $r(t_i, i_i) = 0$ otherwise.

The DNN architecture consists of two branches, one for hashtags and one for images. Each branch has an embedding layer of 64 components. This layer "encodes" the inputs to a feature vector, i.e. it uses the input as an index to obtain the corresponding feature vector. In each iteration, the weights of these vectors are adjusted obtaining at the end of the training a *NxM* matrix where *N* is the number of elements to encode and *M* the number of embedding components. Both branches are joined in a Dot layer that computes the dot product of the previous outputs. The next layers are a Reshape layer to resize the input to a one-dimensional vector and the last layer is a fully-connected layer. The hashtags latent space is formed by the weights of the hashtags embedding layer. An overview of this architecture is shown in Fig. 5.

Finally, the semantic clusters of the hashtags are obtained using the k-means algorithm. To determine the best number of clusters, the elbow method has been applied using the inertia (the sum of squared distances of samples to their closest cluster center) as metric.

*C. **Revise and Retain**: Adding the New Tagged Images to Case Memory*

In the last two phases of the CBR, the solutions are reviewed and the cases where the proposal was incorrect are stored. In a recommendation problem, in the review phase, the proposed tags are compared with those that the user actually used, so unlike most CBRs, no manual review is necessary. Those tags that the user finally used and the system did not recommend are stored in the retain phase. This allows our proposed system to do two things: on the one hand to learn from the tagging habits of the users, which in the case of folksonomies is changing; and on the other hand to add new tags to the system, which in the case of folksonomies the label space is very varied and divergent. Therefore, it is important that each time a new tag is added to the system, both the embeddings of the tags and their semantic clusters are recalculated.

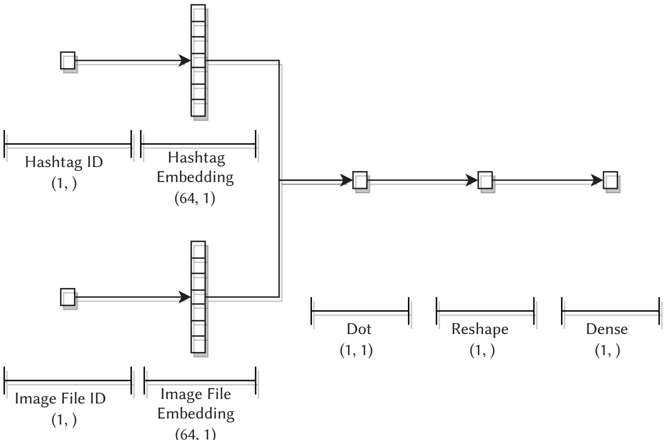

Fig. 5. Overview of the Deep Neural Network architecture for hashtag and image file embeddings.

V. Instagram as a Case Study

Since Instagram is currently the most relevant social network in the marketing field and that it is one of the most used and fastest growing social networks in recent years [1], this work validates our proposal with data obtained from a set of profiles of this social network.

Specifically, the dataset used in this case study arises from [36]. In this work, image and text are both used to categorize different Instagram influencer profiles according to their topics of interest. To this end, a dataset consisting of 10,180,500 posts from 33,935 Instagram influencer profiles is compiled. For each post, the dataset contains image files, captions, hashtags, usertags, number of likes, associated comments and other meta-data. Since it contains all the information necessary for the implementation of our proposal, this dataset is used in our work. In this case, the most relevant information of each post is related to the meta-data of the images and the associated hashtags. In besides applying a novel approach to this data, our goal is not to categorize user posts but to obtain and recommend a set of tags in order to optimize the dissemination of the image posted on Instagram.

In order to obtain the dataset on which to perform the tests, an undersampling of the original dataset was first performed. From the total number of posts we have randomly chosen a 5% sample, equally distributed for each of the topics. From that 5% we extracted the hashtags of each post, resulting in a total of 4,008,534 records. Many of the hashtags found in these records are hardly representative, so a filtering has been performed to eliminate those hashtags that have no more than 0.1% of representativeness. In this way, a final hashtag space of 2,083 hashtags was obtained, ranging from the most representative hashtag #ootd with 30,378 records to #ocblogger with 379 records. However, many hashtags are extensions of others, for example #recipevideo of #recipe. In order not to discard these extensions, they have been grouped under the root hashtag, i.e. for the above case #recipevideo has been replaced by #recipe. In Fig. 6 we can see the tag cloud with the representativeness of each hashtag. The final number of records in the dataset is 3,361,766. This dataset is divided into a training dataset and a test dataset in a ratio of 80/20.

As described in the previous section, the Case-Base of our CBR is constructed from the Case-Representations of the initial cases. The Case-Representations are embeddings of the images obtained from

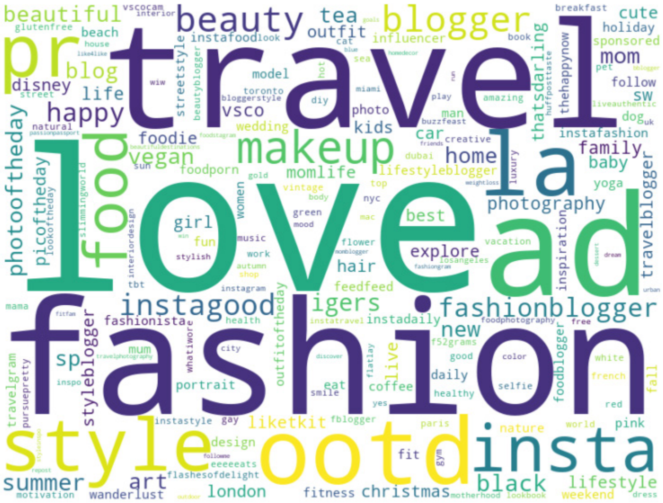

Fig. 6. Influencers dataset hashtags wordcloud.

the CNN trained on the Siamese Neural Network. To train the Siamese Neural Network, a dataset of triples (anchor, positive, negative) has been constructed from the training dataset. The triplets have been created by taking the images as anchor and all those other images that use the same hashtags have been taken as positive, and those others that do not use any anchor hashtags have been taken as negative. That is, the relationship of the triplet (anchor, positive, negative) is whether or not the same hashtags are used. The Fig. 7 shows an example of image triplet.

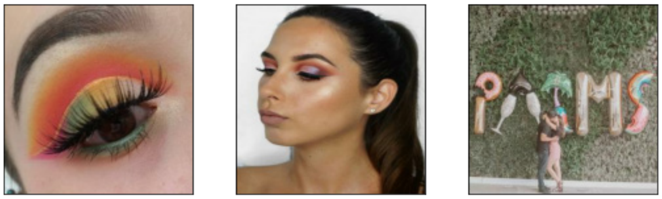

Fig. 7. Sample of image triplet CNN input. The first image is the anchor, the second image is the positive and the third image is the negative.

To obtain the latent space of hashtags, records (x_i, i_i, y_i) have been created where yi indicates whether the hashtag and the image are related $(y_i = 1)$ or not $(y_i = 0)$. This value is taken by the hashtag embedding DNN as a target. In the training dataset there are only positive relationships, so it has been necessary to create negative relationships, i.e. the relationship value is 0. For this purpose, all combinations of pairs pxi, iiq have been taken and those where there is no relationship have been taken as negative.

From this dataset, the latent space of hashtags was trained and obtained. An example of the hashtags closest to the hashtag #ootd in that space can be seen in Table I.

TABLE I. Top 10 Nearest Hashtags to Hashtag #OOTD From Hashtag Latent Space

Hashtag	Distance
#ootd	0.000000
#fashion	0.630845
#fashionblogger	1.063766
#style	1.126536
#outfitoftheday	1.863068
#outfit	1.884869
#styleblogger	2.210241
#instafashion	2.504243
#fashionista	2.607178
#instastyle	2.828929

Finally, the semantic clusters of the hashtags have been obtained from the latent space of hashtags. The algorithm used to obtain the clusters was k-means, using the elbow method to infer the best number of clusters. The results of the elbow method using inertia as a metric can be seen in Fig. 8 In the tests performed, for all possible values of k from 2 to 100, the best possible value was 14 clusters. In Fig. 9 a 2D projection of the latent space of hastags is shown, where all hashtags can be seen colored in red and colored in blue the hashtags closest to #ootd. The 2D projection of the latent space with the hashtags colored per cluster can be seen in Fig. 10.

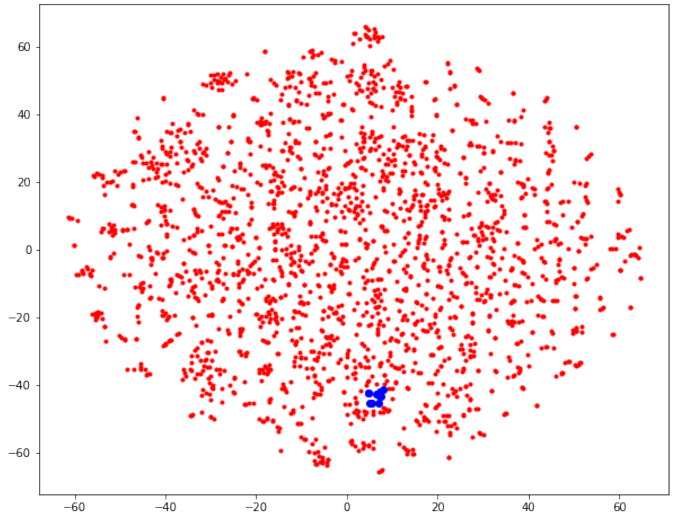

Fig. 9. 2D representation of the hashtag latent space using TSNE. Blue dots are the top 10 nearest hashtags to the hashtag #ootd.

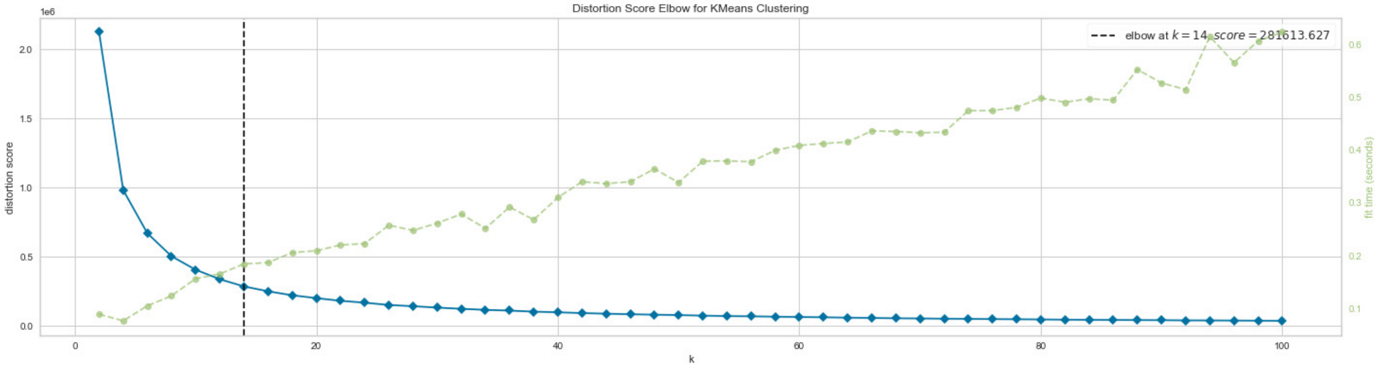

Fig. 8. Results of the elbow method for the number semantic clusters of hashtags.

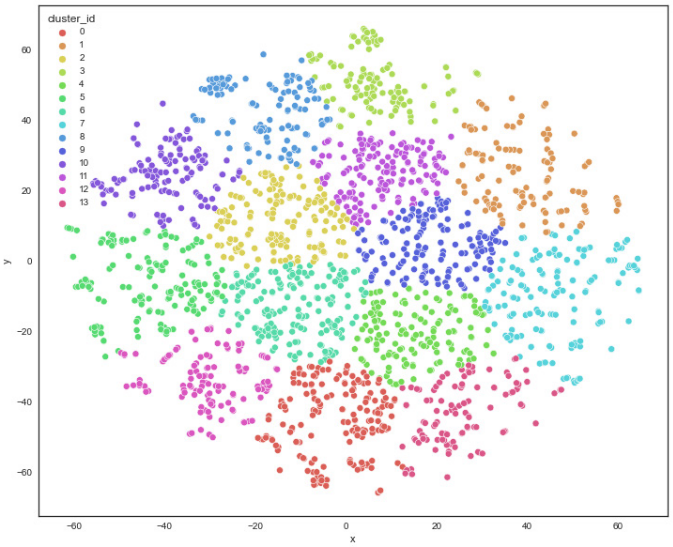

Fig. 10 2D representation of the semantic clusters of hashtags using TSNE.

VI. Results and Discussion

In this section we present the results of the tests performed. To compare our proposal we have chosen two other well-recognized methods also applied for feature extraction such as the color histogram and the Bag of Visual Words (BoVW). These two methods will generate Case-Representations against which to compare our proposal. All three methods use the same Case-Base filled from the Case-Representations of the training dataset. In the tests, a prediction of 10 hashtags has been performed for each new case. After several initial tests, the optimal value of nearest neighbors for the retrieve phase is 1000. In addition, since the number of cases in the Case-Base is initially very large and penalizes the query times in the retrieval phase, it has been reduced using Random Selection Undersampling method, reducing the cases taking into account the minority hashtag. This reduction of cases has not penalized the results and has improved the query times.

Unlike the metrics commonly used in recommender systems such as MAE and RMSE, we have chosen a different set of metrics as we find they are more suited to the purpose of the proposed system. The metrics chosen to evaluate the models were precision, recall and f1-score. Additionally, the distance between the gravitational center of the proposal's embeddings and the gravitational center of the user's hashtags' embeddings has been calculated. In this way, we can compare the quality of the model in terms of the semantic quality of the proposal.

In our tests we evaluate the overall matching of the proposed hashtags with the user's hashtags, the matching of at least one of the proposed hashtags with the user's hashtags, and the matching of the semantic clusters of the proposal with the semantic clusters of the user's hashtags.

TABLE II. Results for Full Hashtags Recommendation Match

Method	Precision	Recall	F1	Dst
Color hist.	0.0421	0.0602	0.0444	39.6797
BoVW	0.0479	0.0580	0.0439	38.8396
Proposal	0.0629	0.0696	0.0502	38.6142

The results of the first test can be seen in Table II. As we can see, our system improves the other two, both in precision and recall (and therefore in f1-score). This shows that our system not only makes fewer errors, but also hits more user hashtags. Moreover, the distance between the hashtags is smaller than in the other two, i.e., even if our system makes a mistake in the hashtag to recommend, it is semantically closer than the recommendations of the other two systems.

In the case of getting at least one of the recommended hashtags right, our proposal improves on the other two, as shown in Table III. As before, our model makes fewer errors and covers more right hashtags than the other two models. It is interesting to highlight that the color histogram in this case improves the BoVW, unlike the previous case. As a result, although BoVW has a higher precision and recall in the first case, it should do so in fewer recommendations than the color histogram. In other words, it has better metrics but does good recommendations in fewer cases.

TABLE III. Results for at Least One Hashtag Recommendation Match

Method	Precision	Recall	F1
Color hist.	0.1266	0.2044	0.1564
BoVW	0.1240	0.1451	0.1337
Proposal	0.1540	0.2216	0.1817

Finally, in the case where the user's recommendations and hashtags belong to the same semantic clusters, our system also outperforms the others (Table IV). In this case it reaches 22% precision and 33% recall, i.e., approximately one out of four recommendations matches semantically and the recommendations cover one third of the topics (within the folksonomy) that the user wants to tag in the image.

TABLE IV. Results of Hashtags Belonging to the Same Semantic Groups of Hashtags

Method	Precision	Recall	F1
Color hist.	0.2021	0.2776	0.2232
BoVW	0.2057	0.2733	0.2248
Proposal	0.2247	0.3269	0.2593

Although the precision and recall are low in general, we must keep in mind that we are evaluating a recommendation problem and that these values are usually low. We are not evaluating the exact prediction but how good the recommendations are. It should be taken into account that aspects such as the influence of the recommendation on the user when choosing hashtags cannot be evaluated a priori.

The results of the research experiments carried out in this article can be found at this link.

VII. Conclusions

This proposal has presented an intelligent system to suggest tags for an image previously submitted to social networks. Instagram tags are recommended based on the image features and previous experiences on other similar uploaded posts. Therefore, a CBR architecture that learns from previous solutions is applied. As a first step, the system is populated with tagged images submitted to the social network. Then, the system compares a new image manually selected with similar images stored in their memory. Finally, the recommendation of the system is a set of tags which helps to disseminate an image in a social network. Thus, this work addresses a multi-label problem.

In order to demonstrate the validity of the system and its independence of the dataset, a cross validation was carried out in order to evaluate the of the system of new tags.The overall results of the experiments carried out emphasize that the system can suggest concrete words as tags that influences in the visibility of the post. Another important point regarding the results is that, although the technique selected in our proposal to obtain the case representation is a neural network, the color histogram and the SIFT attributes grouped as BoVW could also be valid.

The use of folksonomies is a human-machine collaborative approach. Tags are automatically obtained from the image properties of those obtained for similar cases, but in addition, the tagging behavior of Instagram users is taken into account, so the algorithm adapts to new trends and randomness in the tagging process is reduced.

Words suggested as tags are always based on previous cases, so the system does not infer new knowledge based on the semantics of the words. For a future work, we propose the implementation of Natural Language Processing techniques in order to predict new words based on previous cases and the involvement of Instagram accounts to test the recommender in the social network.

References

[1] Statista, "Number of monthly active instagram users from january 2013 to december 2021," 2022. [Online]. Available: https:

[2] //www.statista.com/statistics/253577/ number-of-monthly-active-instagram-users/.

[3] S. Sood, S. Owsley, K. J. Hammond, L. Birnbaum, "Tagassist: Automatic tag suggestion for blog posts.," in *ICWSM*, 2007.

[4] M. Sohn, S. Jeong, J. Kim, H. J. Lee, "Augmented context-based recommendation service framework using knowledge over the linked open data cloud," *Pervasive and Mobile Computing*, vol. 24, pp. 166–178, 2015.

[5] W. Chen, Z. Li, "A study of tag-based recipe recommendations for users in different age groups," in *International Symposium on Emerging Technologies for Education*, 2016, pp. 315–325, Springer.

[6] D. Eck, P. Lamere, T. Bertin-Mahieux, S. Green, "Automatic generation of social tags for music recommendation," in *Advances in neural information processing systems*, 2008, pp. 385–392.

[7] O. Celma, "Music recommendation," in *Music recommendation and discovery*, Springer, 2010, pp. 43– 85.

[8] J.-H. Su, H.-H. Yeh, S. Y. Philip, V. S. Tseng, "Music recommendation using content and context information mining," *IEEE Intelligent Systems*, vol. 25, no. 1, 2010.

[9] A. Van den Oord, S. Dieleman, B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643– 2651.

[10] X. Wang, Y. Wang, "Improving content-based and hybrid music recommendation using deep learning,"in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 627–636, ACM.

[11] Y. Choi, J. Kim, E. Yun, S. Lee, D. Kim, "A new image search and retrieval system using text and visual features," in *WebNet World Conference on the WWW and Internet*, 2000, pp. 742–743, Association for the Advancement of Computing in Education (AACE).

[12] J.-W. Huang, C.-Y. Tseng, M.-C. Chen, M.-S. Chen, "Pisar: Progressive image search and recommendation system by auto-interpretation and user behavior," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, 2011, pp. 1442–1447, IEEE.

[13] M. De Gemmis, P. Lops, G. Semeraro, P. Basile, "Integrating tags in a semantic content-based recommender," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 163–170, ACM.

[14] P. Lopez-de Arenosa, B. Díaz-Agudo, J. A. Recio- García, "Cbr tagging of emotions from facial expressions," in *International Conference on Case-Based Reasoning*, 2014, pp. 245–259, Springer.

[15] S. Craw, B. Horsburgh, S. Massie, "Music recommendation: audio neighbourhoods to discover music in the long tail," in *International Conference on Case-Based Reasoning*, 2015, pp. 73–87, Springer.

[16] S. Nasiri, J. Zenkert, M. Fathi, "A medical case-based reasoning approach using image classification and text information for recommendation," in *International Work-Conference on Artificial Neural Networks*, 2015, pp. 43–55, Springer.

[17] M. B. Chawki, E. Nauer, N. Jay, J. Lieber, "Tetra: A case- based decision support system for assisting nuclear physicians with image interpretation," in *International Conference on Case-Based Reasoning*, 2017, pp. 108–122, Springer.

[18] D. López-Sánchez, J. M. Corchado, A. G. Arrieta, "A cbr system for image-based webpage classification: Case representation with convolutional neural networks," 2017.

[19] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human- computer joint exploration," *ACM Computing Surveys (CSUR)*, vol. 44, no. 4, pp. 1–24, 2012.

[20] J. Li, J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 6, pp. 985–1002, 2008.

[21] X. Li, C. G. Snoek, M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 180–187.

[22] H. T. Nguyen, M. Wistuba, L. Schmidt-Thieme, "Personalized tag recommendation for images using deep transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 705–720, Springer.

[23] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, G. Stumme, "Tag recommendations in folksonomies," in *European conference on principles of data mining and knowledge discovery*, 2007, pp. 506–514, Springer.

[24] A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, "Folkrank: A ranking algorithm for folksonomies," 2006.

[25] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, G. Thallinger, "Automatic image annotation using visual content and folksonomies," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 97–113, 2009.

[26] S. Lee, W. De Neve, K. N. Plataniotis, Y. M. Ro, "Map-based image tag recommendation using a visual folksonomy," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 976–982, 2010.

[27] E. Amador-Domínguez, E. Serrano, D. Manrique, J. Bajo, "A case-based reasoning model powered by deep learning for radiology report recommendation," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 2, 2021.

[28] M. Benamina, B. Atmani, S. Benbelkacem, "Diabetes diagnosis by case-based reasoning and fuzzy logic," *IJIMAI*, vol. 5, no. 3, pp. 72–80, 2018.

[29] O. R. Zaíane, "Building a recommender agent for e- learning systems," in *Computers in education, 2002. proceedings. international conference on*, 2002, pp. 55–59, IEEE.

[30] D. G. Lowe, "Distinctive image features from scale- invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[31] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "Speeded- up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[32] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*, 2011, pp. 2564–2571, IEEE.

[33] J. Yang, Y.-G. Jiang, A. G. Hauptmann, C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 197–206, ACM.

[34] R. Chakravarti, X. Meng, "A study of color histogram based image retrieval," in *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, 2009, pp. 1323–1328, IEEE.

[35] M. Hassan, C. Bhagvati, "Evaluation of image quality assessment metrics: Color quantization noise," *Evaluation*, vol. 9, no. 1, 2015.

[36] L. Torrey, J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[37] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, W. Wang, "Multimodal post attentive profiling for influencer marketing," in *Proceedings of The Web Conference 2020*, 2020, pp. 2878–2884.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.K. Weiss, T. M. Khoshgoftaar, D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1– 40, 2016.

[39] E. Hoffer, N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*, 2015, pp. 84–92, Springer.

[40] I. Melekhov, J. Kannala, E. Rahtu, "Siamese network features for image matching," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 378–383, 2016.

Lucía Martín-Gómez

Lucía Martín Gómez has a PhD in Computer Engineering, an official master's degree in Intelligent Systems and a degree in Computer Engineering from the University of Salamanca. In the research context, she has made several scientific publications and has collaborated as organizing committee of some international conferences framed in multiple areas within artificial intelligence. She has participated in several research projects related to IoT, social network analysis, big data and Industry 4.0 at national and European level. In 2018 she was granted a Pre-doctoral Scholarship for research workers training awarded by the Junta de Castilla y León (España). She has developed her professional career working as a data scientist in big data projects, and is currently a Professor at the Pontifical University of Salamanca.

Javier Pérez-Marcos

Javier Pérez has a degree in Computer Engineering at the University of Salamanca, a master's degree in Intelligent Systems at the University of Salamanca, and is currently a PhD candidate at the University of Salamanca. He worked for two years as Researcher at BISITE research group at the University of Salamanca, and for three years as Data Scientist and Big Data Manager at Smart Internet Internacional Sl. He is currently working as Data Engineer at Frogtek Sl. In addition, he has been Visiting Professor at the University of Salamanca, Visiting Professor at the University of Deusto and for the last three years Associate Professor at the Pontifical University of Salamanca.

Rebeca Cordero Gutiérrez

Rebeca Cordero Gutiérrez holds a PhD in Logic and Philosophy of Science with a specialization in Social Studies of Science and Technology from the University of Salamanca. Her doctoral thesis was developed in the field of horizontal social networks and their business and social impact. She has been a university lecturer for more than 10 years and has participated as research staff in several competitive projects. Her lines of research focus on the impact of ICT and social networks in business and education. She has taught courses and seminars related to new technologies and business management. She is the author of several articles in relevant journals, and has participated in numerous national and international conferences.

Daniel Hernández de la Iglesia

Daniel Hernández de la Iglesia is a Ph.D in Computer Engineering, a Technical Engineer in Computer Systems and a Degree in Computer Engineering from the University of Salamanca. He has an official master's degree in Intelligent Systems and in recent years has been linked to different research groups in the field of artificial intelligence where he has participated in dozens of national and international research projects. He is the author of different book chapters and has presented more than twenty research papers in different international congresses. In addition, it has numerous scientific publications in international impact journals indexed in the JCR reference ranking. He has been awarded the first prize of the Open Data Contest organized by the Junta de Castilla y León (Spain) in 2013, and with the first innovation award for the best research project awarded by the Junta de Castilla y León (Spain) in 2016. He is currently a Professor at the Pontifical University of Salamanca.