




# Automatic assignment of microgenres to movies using a word embedding-based approach

Carlos González-Santos<sup>1</sup> · Miguel A. Vega-Rodríguez<sup>2</sup>  · Joaquín M. López-Muñoz<sup>3</sup> · Iñaki Martínez-Sarriegui<sup>3,4</sup> · Carlos J. Pérez<sup>1</sup>

Received: 8 March 2022 / Revised: 10 September 2023 / Accepted: 3 October 2023 /  
Published online: 3 November 2023  
© The Author(s) 2023

## Abstract

Streaming services are increasingly leveraging Artificial Intelligence (AI) technologies for improved content cataloging, user experiences in content discovery, and personalization. A significant challenge in this domain is the automated assignment of microgenres to movies. This study introduces and evaluates approaches based on clustering, topic modeling, and word embedding to address this task. The evaluation employs a preprocessed dataset containing movie-related data—title tags, synopses, genres, and reviews—alongside a predefined microgenre list. Comparisons of three activation functions (binary step, ramp, and sigmoid) gauge their effectiveness in augmenting microgenre tags. Results demonstrate the superiority of the word embedding approach over clustering and topic modeling in terms of mean accuracy. Even more, the word embedding approach stands as the sole fully automated solution. Analysis indicates that incorporating review-based tags introduces noise and undermines accuracy. Besides, the word embedding approach yields optimal outcomes using the sigmoid function, effectively doubling assigned tags while maintaining matching quality. This sheds light on the potential of word embedding methods within the movie domain.

**Keywords** Movie microgenre · Word embedding · Semantic similarity · Clustering · Topic modeling · Activation function

## 1 Introduction

The volume of information in the Internet is growing exponentially and it is stored in diverse formats such as text, image, audio, video, etc. Since these information sources contain large data collections, their overwhelming quantity makes some processes difficult, such as content discovery, where users want instantaneous answers. The use of tags can speed up these processes. Tags are descriptive labels of one or a few words associated with a particular piece of content [1]. They help to obtain content with semantic sense and actionable information. Unlike structured data, tags are usually open-ended and aim to capture implicit information

---

✉ Miguel A. Vega-Rodríguez  
mavega@unex.es

Extended author information available on the last page of the article

not readily available without an analysis of the content itself. They are useful in many fields such as information retrieval [2], recommender systems [3], or sentiment analysis [4], among others.

More recently, tags have also become popular in streaming services as their content catalogues grow in size and complexity [5]. In the movie industry, each company uses its own sets of tags based on heterogeneous typologies. Among these tags, thematic tags map the content into a relatively small number of categories based on its plot subject and artistic qualities, i.e., tags describing a limited classification of genres (drama, action, comedy, etc.) [6].

Microgenres, initially popularized by Netflix<sup>1</sup> but widespread nowadays, go beyond traditional genres and propose a finer classification of content, so that entries falling under the same microgenre are much more closely related from a thematic viewpoint [7]. Some examples of microgenres may be: ‘absurdist humor’, ‘animation for adults’, ‘classic Hollywood musical’, ‘dysfunctional family’, ‘natural disasters’, or ‘classic noir novel’. In fact, microgenres can be considered as a kind of thematic tags. The use of appropriate thematic tags to create more concrete and diverse genres (microgenres) instead of conventional genre tags is an open problem. Microgenres, with a broader and more specific theme coverage, enable a more detailed and enriched categorization of movies. However, addressing this challenge requires a substantial editorial effort or some novel approaches that could automatically perform this task providing accurate results.

To the best of the authors’ knowledge, up to now, no method has been developed to automatically build microgenres from tags with the objective of assigning microgenres to movies. Several investigations have been reported related to this novel problem, not exactly for microgenres, but for genres or tagging, which are less complex problems. Wu et al. [8] exploited user reviews to perform an automatic tagging of movies. Regarding the movie genre classification, Guehria et al. [9] used movie synopses and a document-based embedding approach called Doc2vec. Yu et al. [10] classified movies by using neural networks and their trailers. Kundalia et al. [11] performed this genre classification by using movie posters and neural networks employing knowledge transfer learning. Finally, Mangolin et al. [12] mixed all previous resources to develop a multimodal approach for genre classification.

In this research, the problem of automatic assignment of microgenres to movies is addressed. This automated process works with tags previously extracted from the information about the movie. Some information providers and online communities like IMDb<sup>2</sup> and MovieLens<sup>3</sup> provide crowdsourced tags for movie content, but current Natural Language Processing (NLP) techniques like automatic keyword extraction [13] are well suited for the automatic generation of these tags. This problem requires a convenient way of matching content entries, which are described by automatically generated tag sets, to a predefined list of microgenre names. Previously, these microgenres have been semantically defined by matching each of them with the most semantically related tags. Thus, a word embedding-based approach has been designed and implemented for this purpose. Experiments have been carried out with a movie dataset specifically developed for this research and comparisons have been performed with other methods adapted to solve this problem. The main contributions of this work are:

- The problem of automatic assignment of microgenres to movies is addressed from a scientific viewpoint for the first time.

---

<sup>1</sup> <https://www.netflix.com/>

<sup>2</sup> <https://www.imdb.com/>

<sup>3</sup> <https://movielens.org/>

- Three techniques (clustering, topic modeling, and word embedding) have been specifically designed, implemented, and applied to solve this assignment problem.
- A dataset has been built within the movie domain to evaluate the approaches implemented for this problem.
- The behavior of review tags has been analyzed verifying that they introduce noise into the semantic definition of the microgenres, represented by these tags, and, consequently, the microgenre assignment worsens.
- Three activation functions (binary step, ramp, and sigmoid) have been implemented and compared with the objective of increasing the number of tags assigned to microgenres.

Once the underlying problem has been presented in this section, the rest of the article is organized as follows. In Section 2, the followed methodology and proposed approaches are detailed. The experimental settings, evaluation metrics, results, their discussion and comparisons are included in Section 3. Finally, Section 4 shows the conclusions and future research.

## 2 Methodology

In this section, the used datasets and their preprocessing are presented. Then, three approaches based on clustering, topic modeling, and word embedding (using Word2vec [14]) are proposed to solve this problem. Finally, the three considered activation functions will be described.

### 2.1 Datasets

In this research, two datasets are involved. The first one is a movie collection and the second one a microgenre collection. The movie collection is part of a database provided by Optiva Media<sup>4</sup> and enriched and curated by Metadatol. This database consists of a collection of 3,413 movies and 82,297 tags, which provide information about the movies. Most of the tags have been extracted from NLP analysis of metadata obtained from popular and publicly available sources, like IMDb, Rotten Tomatoes<sup>5</sup>, and FilmAffinity<sup>6</sup>. The rest of tags comes from an external taxonomy specifically developed to enrich the tag set.

The tags are associated with movies and are divided into 3 categories: descriptions (41,003 tags extracted from the movie synopses), reviews (41,111 tags extracted from users' comments from the previously mentioned platforms), and genres (183 tags).

The second dataset consists of a list of movie microgenres. This list has been extracted from Filmin<sup>7</sup>, a popular cinematographic platform. Its website has a public list of 1,233 topics or microgenres in which movies are categorized, like 'Plane accident', 'Declaration of love', or 'Assassinated politician'.

<sup>4</sup> <https://www.optivamedia.com/>

<sup>5</sup> <https://www.rottentomatoes.com/>

<sup>6</sup> <https://www.filmaffinity.com/>

<sup>7</sup> <https://www.filmin.es/>

## 2.2 Preprocessing

Datasets for text processing tasks usually contain words that do not provide any valid information. In order to optimize the model performance, the tag set has been preprocessed through the following steps:

1. Digits and symbols removal. Word2vec works with words formed only by letters to provide results with a reliable semantic level. Then, all digits and symbols, such as punctuation marks, ampersands or hashes, are removed.
2. Stopwords removal. There are many words that appear frequently and they do not offer any semantic value. These words (articles, prepositions, etc) are removed from the tag collection. The tags that only contain stopwords are removed from the tag collection, e.g. ‘All By Myself’. Concretely, the set of English stopwords corresponding to the NLTK<sup>8</sup> package has been used.
3. Duplicates reduction. Tags that appear several times after applying the two previous steps are reduced to one instance, whereas their assignation to movies is kept.

After applying the preprocessing steps, the new tag set has 75,310 tags, divided into 37,321 description tags, 37,867 review tags, and 122 genre tags.

The microgenre collection has also several preprocessing steps to improve its semantic content:

1. Translation. Microgenres were extracted in Spanish, so they have been translated into English.
2. Stopwords removal. Analogously to the tag case, words with no semantic value are removed from the microgenre collection.
3. Specific microgenres removal. Some microgenres that describe opinions, names of people or locations may introduce noise due to its semantic meaning, so they have been removed. For example, ‘Stephen King’, where the surname can be confused with the monarch figure.
4. Microgenres adaptation. Some words that appear in many microgenres introduce noise and reduce semantic differences among them. Then, microgenres like ‘Horror Cinema’ and ‘Reggae Music’ remain as ‘Horror’ and ‘Reggae’.

After applying the previous preprocessing steps, the microgenre collection contains 666 microgenres. Hereafter, the two preprocessed datasets will be used.

## 2.3 Techniques matching microgenres

Given the previous databases, the objective is to assign microgenres to movies. The following notation will be considered through the article. Let  $M = \{M_s\}_{s=1}^S$  be a collection of  $S$  movies. Each movie  $M_s$  is represented by a vector with  $N_{M_s}$  tags  $t_{M_s} = (t_{M_s,1}, \dots, t_{M_s,N_{M_s}})$ . Each tag is classified into one or more of the  $T$  microgenres  $g_1, \dots, g_T$ . In order to determine the most related microgenres to each movie, three approaches are proposed.

### 2.3.1 Clustering

Clustering is a very used unsupervised classification technique [15]. It divides a collection of data points into groups (clusters) by similarity. The main types of clustering models are

<sup>8</sup> <https://www.nltk.org>

partitional and hierarchical models. Partitional models assign data points into a number of clusters by optimizing some specific function, whereas hierarchical models build clusters by dividing the found patterns in many iterations, providing a hierarchical structure. K-Means [16] (partitional model) and BIRCH [17] (hierarchical model) are two of the most used methods from these two clustering types.

The assignment problem is solved in two steps for K-Means and BIRCH. Firstly, a classification of the tags into microgenres is performed with the corresponding clustering. Then, a matching score is calculated for every movie. Clustering has as input a matrix that contains tags as rows and movies as columns. The matrix stores the assignation of tags to movies. Then, the clustering is performed. The output is a set of clusters with tags. Clusters are labeled only by a number (Cluster 1, Cluster 2, Cluster 3, ...), so next step is to identify and match them with the microgenre collection. This process is done manually by three experts, who match a microgenre to a cluster by checking those tags that compose each cluster. In the case of K-Means, tags are ordered by distance to the centroid, which makes the association task easier, whereas BIRCH does not order tags, so each one is considered within its cluster with the same relevance.

Once tags have been classified into microgenres, the matching scoring is performed. The input is the tag-movie matrix, the tag clusters, and the movie collection. The frequency of appearances in movies is obtained for every tag within a microgenre. Then, for every movie, the mean appearance frequency of each microgenre in the movie is calculated by averaging the appearance frequencies of the tags from each movie with respect to that microgenre. Finally, all microgenres (assigned to each movie) are sorted by their mean appearance frequency in descending order as the final output.

### 2.3.2 Topic modeling

Topic modeling is a very useful technique in the field of text mining [18]. The objective of topic modeling is to discover topics that are able to represent a document collection, that is, to discover what the topics might be and what each document's balance of topics is. Latent Dirichlet Allocation (LDA) is one of the most popular topic models [19].

The proposed approach for topic modeling has also two steps. In the first step, the input of LDA is a collection of documents, each one containing all the tags of each movie. As LDA takes every word as an entity, each tag composed by more than one word has used underscores ('\_') to join its words. Then, LDA is executed and the assignation of tags to topics is obtained for every movie, in addition to a list with the most representative tags for each topic. Three experts used these lists to match manually the microgenres with each of the topics.

In the second step, the matching scoring uses as input the classification of tags into microgenres (topics), the tag-movie matrix, and the movie collection. This process is analogous to the one implemented for clustering (explained in the previous subsection).

### 2.3.3 Word embedding

Word embedding is a recent branch in the field of language modeling [20]. It builds distributed word vectors of fixed length that allow to perform vector operations with a word collection. Word embedding is a well-known method to solve natural language processing tasks [21].

Word2vec<sup>9</sup> is a popular implementation of word embedding based on very large datasets with the aim of establishing semantic relations between words or phrases [14]. The semantic similarity is calculated through the cosine distance between the two vectors representing those two words/phrases. Word2vec is used to solve problems like document similarity detection [22], or keyword extraction [23], among others. A pretrained vector model has been used for this research. It has been trained on the Google News dataset, which contains about 100 billion words [24]. The model has 300-dimensional vectors for 3 million words and phrases. It provides very efficient results in terms of semantics with vector operations such as  $vec('King') - vec('Man') + vec('Woman') = vec('Queen')$ .

This Word2vec-based approach is fully automatic, which represents a clear advantage with respect to the previous approaches. Word2vec uses as input the tag collection and the list of microgenres. For every movie, the model performs an automatized process that takes a tag and calculates its semantic similarity with respect to each microgenre, i.e., a numeric value in the range [0,1]. If a tag or a microgenre has more than one word, Word2vec sums each one of the word vectors before it calculates the semantic similarity. Then, the tag is assigned to those microgenres where the semantic similarity is greater or equal than a threshold,  $lim_{step}$ , which takes a fixed value within the range [0,1]. Following this process, there exists the possibility that a tag can be assigned to multiple microgenres or even to none of them.

After the tag classification, the matching scoring is performed. The classification of tags into microgenres, the tag-movie matrix, and the movie collection are taken as input. For every movie, the mean semantic similarity of every microgenre with the movie is calculated by averaging the semantic similarity of the tags from each movie with respect to that microgenre. Finally, for each movie, all the assigned microgenres are sorted by their mean semantic similarity in descending order as the final output.

## 2.4 Assigning tags to microgenres through activation functions

The Word2vec-based approach follows a strict requirement (threshold) with respect to the assignation of a tag to a microgenre. This requirement can be made more flexible through activation functions. An activation function describes through fuzzy logic the grade in which a data point belongs to a defined set. Three activation functions (binary step, ramp, and sigmoid) have been applied to the Word2vec-based approach proposed here to determine which tags are assigned to each microgenre.

The Word2vec-based approach implements the binary step function, which is involved in the tag classification task. The binary step function takes an input value and returns either 0 or 1, depending on whether the input value is below or above a certain threshold. Regarding the Word2vec-based approach, the activation function uses the semantic similarity between a tag and a microgenre to decide whether the tag belongs to that microgenre. As previously stated, the parameter  $lim_{step}$  determines which tags are assigned to each microgenre. If the semantic similarity is greater or equal than  $lim_{step}$ , the tag is assigned to that microgenre (i.e., the function returns 1). If the semantic similarity is below the threshold, the tag is not assigned to the microgenre (i.e., the function returns 0). The binary step function is represented as:

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < lim_{step} \\ 1 & \text{if } lim_{step} \leq x \leq 1 \end{cases} . \quad (1)$$

<sup>9</sup> <https://code.google.com/archive/p/word2vec/>

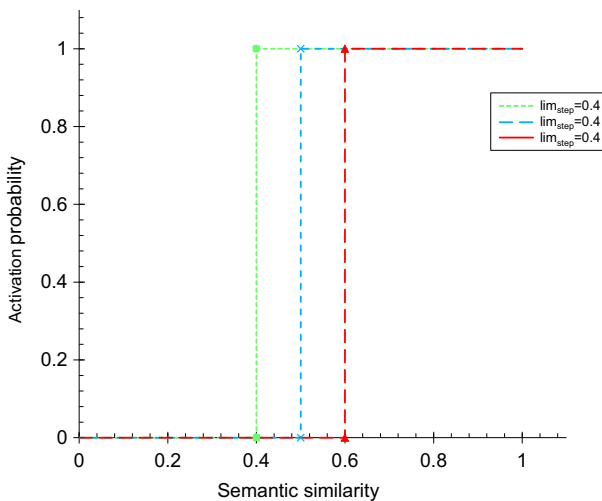
Figure 1 displays the three configurations of the binary step function with parameter  $lim_{step} = 0.4, 0.5, 0.6$ . The lower the value of  $lim_{step}$ , the higher the number of tags that are assigned to a microgenre, which may introduce noise into the results due to the lack of semantic similarity.

The ramp and sigmoid functions perform the tag classification on a different, more flexible, way. In this process, the semantic similarity is considered as an input to calculate the activation probability following one of the two functions. Then, a random number in the range  $[0,1]$  is generated. If the activation probability is greater or equal than the random number, the tag is assigned to that microgenre with its semantic similarity.

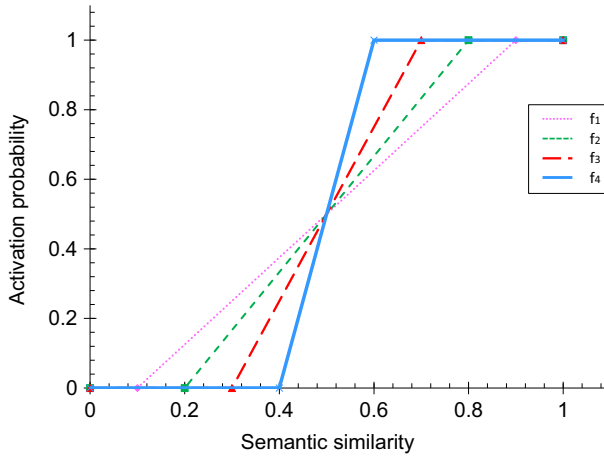
The ramp function is defined as a continuous function that increases linearly from zero to one over a specified interval, i.e., a lower limit and an upper limit, keeping constant outside the interval. If the semantic similarity is less than the lower limit, the activation probability is 0. If the semantic similarity is greater or equal than the upper limit, the activation probability is 1. Otherwise, the activation probability is calculated by a function defined by the straight line that connects the lower and upper limits. Four configurations following a balanced criterion have been considered to select the one providing the best results. The general ramp function for this approach is defined as:

$$f_n(x) = \begin{cases} 0 & \text{if } 0 \leq x < \frac{n}{10} \\ \frac{5}{5-n}x + \frac{12n}{n-5} & \text{if } \frac{n}{10} \leq x < 1 - \frac{n}{10} \\ 1 & \text{if } 1 - \frac{n}{10} \leq x \leq 1 \end{cases} \quad (2)$$

Figure 2 displays the four configurations of the ramp function with parameter  $n = 1, 2, 3, 4$ . The behavior of these functions becomes more flexible than the binary step functions for the assignment of tags to microgenres, allowing for a wider range of similarity values to be considered. The function with  $n = 4$  is the most similar one to the binary step function.



**Fig. 1** Graphical representation of the binary step function using the three configurations for  $lim_{step} = 0.4, 0.5, 0.6$



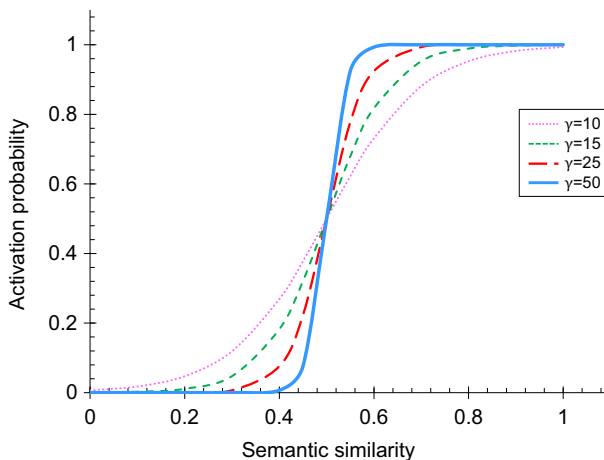
**Fig. 2** Graphical representation of the ramp function using the four configurations for  $f_n, n = 1, 2, 3, 4$

The sigmoid function is the last considered activation function and it takes values in the range  $[0,1]$ . It has a unique parameter  $\gamma$ , which adjusts the slope of the function, i.e.:

$$f(x) = \frac{1}{1 + e^{-\gamma(x-0.5)}}. \tag{3}$$

Figure 3 shows the graphical representation of the sigmoid functions with parameter  $\gamma = 10, 15, 25, 50$ . The function with  $\gamma = 50$  is the most similar one to the binary step function. Regarding the assignment problem, the sigmoid function is the most permissive one, due to the fact that, with some  $\gamma$ , every tag has a non-zero probability of being assigned to each microgenre.

Both ramp and sigmoid functions provide more permissive results at the tag classification task. This leads to microgenres having more associated tags and, consequently, movies are represented by a larger number of tags and microgenres.



**Fig. 3** Graphical representation of the sigmoid function using four configurations with  $\gamma = 10, 15, 25, 50$



### 3 Results

In this section, the experimental settings and evaluation metrics are presented first. Next, the three proposed matching approaches are compared. Then, the impact of review tags is analyzed with the best matching approach. Finally, the performance of the three activation functions is compared.

#### 3.1 Experimental settings

A diverse set of 11 popular movies has been selected to perform the experiments. Then, every approach performs a matching scoring to obtain the 10 most representative microgenres for each movie. As there is no gold standard in this domain, three experts were used in the evaluation task.

Both K-Means and BIRCH models have a parameter, i.e., the number of clusters that has been set to 50. LDA needs to adjust three parameters, i.e., the number of topics and hyperparameters  $\alpha$  and  $\beta$ , that have been set to 50, 50 and 0.01, respectively. The Word2vec-based approach needs to select the activation function parameter, which has been set to  $lim_{step} = 0.5$  for the binary step function,  $n = 4$  for the ramp function, and  $\gamma = 50$  for the sigmoid function.

The preprocessed set of tags for these 11 movies, the reduced list of the 50 microgenres, and the full list of 666 microgenres are provided as supplementary material.

The experiments have been performed on a computer with an Intel Core i5-5200U CPU with 8GB RAM, and Windows 10 as operating system. The matching approaches have been implemented in Python 3.9 with the PyCharm 2021.1.3 IDE, with the exception of LDA, where the Java toolkit MALLET 2.0.8 [25] has been used.

#### 3.2 Evaluation metrics

Owing to the limitation of the data (there is no a gold standard), metrics such as recall and F1 scores are not appropriate. For this reason, the model performance has been assessed by using the following metrics:

- Mean accuracy. The accuracy measures how many of the selected microgenres are correctly assigned to a movie  $M_s$ , i.e.:

$$Accuracy(M_s) = \frac{\text{number of correct microgenres in } M_s}{\text{total number of microgenres in } M_s}. \quad (4)$$

Given  $S$  movies, the mean accuracy is calculated as:

$$Accuracy_{mean} = \frac{\sum_{s=1}^S Accuracy(M_s)}{S}. \quad (5)$$

- Mean linking rate. The linking rate counts the number of tags associated to a movie  $M_s$  that are linked to microgenres, i.e.,  $N_{M_s}$ . This metric measures the capacity of an activation function to obtain the maximum possible amount of information for microgenre description. A larger linking rate results in a more enriched descriptive information used

to define a microgenre, which may provide a better assignment of microgenres to movies. Given  $S$  movies, the mean linking rate is calculated as:

$$Linking_{mean} = \frac{\sum_{s=1}^S N_{M_s}}{S}. \quad (6)$$

### 3.3 Comparison of matching approaches

The first experiment compares the proposed matching approaches. For this experiment, the number of microgenres has been set to 50, due to difficulty of manually matching a large number of 666 microgenres to the clusters and topics generated by the clustering approaches (K-Means and BIRCH) and the topic modeling approach (LDA), respectively. Note that the Word2vec-based approach does not need any manual operation, since it automatically populates each microgenre with its most semantically similar tags.

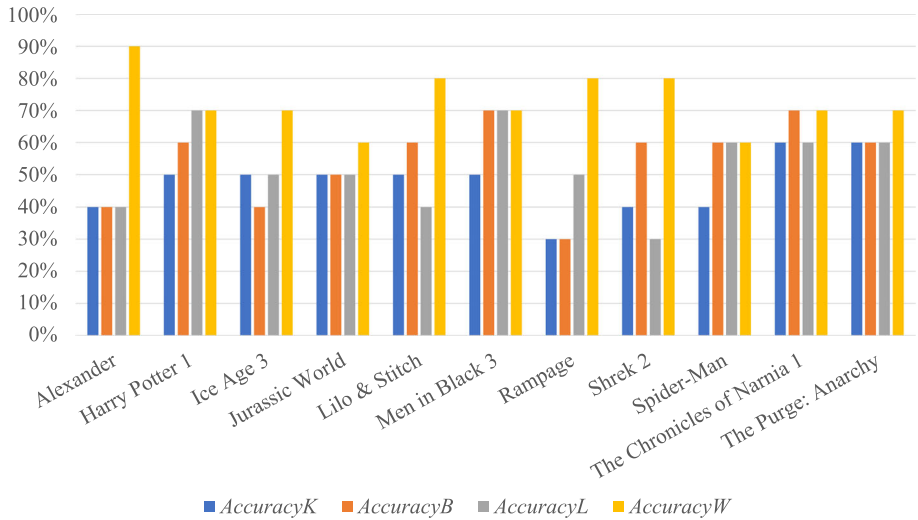
Both K-Means and BIRCH divide the movie collection into 50 clusters (the number of microgenres for this experiment), whereas LDA splits the movie collection into 50 topics. Finally, Word2vec calculates the semantic similarities of every tag with respect to the 50 microgenres under consideration.

Table 1 shows the accuracy results. K-Means, BIRCH, and LDA provided mean accuracies of 47.27%, 54.55%, and 52.73%, respectively, whereas Word2vec achieved a mean accuracy of 72.73%. K-Means, BIRCH, and LDA provided similar accuracies, despite of their different models. Word2vec-based approach, i.e. word embedding, clearly outperforms clustering and topic modeling approaches by increasing the mean accuracy by around 20%. This remarkable improvement in accuracy provides evidence of the specialized effectiveness of the Word2vec-based approach in capturing the textual and semantic relationships among words. In contrast, clustering and topic modeling methods, while versatile in their utility beyond the domain of textual semantics, they do not provide a high performance in this task.

Furthermore, in order to provide a visual representation of these results, Fig. 4 has been used to represent them in a bar chart. It can be observed that Word2vec-based approach

**Table 1** Accuracy and Accuracy<sub>mean</sub> provided by K-Means (K), BIRCH (B), LDA (L), and Word2vec (W) for 50 microgenres and 11 movies

Movie	Accuracy <sub>K</sub>	Accuracy <sub>B</sub>	Accuracy <sub>L</sub>	Accuracy <sub>W</sub>
Alexander	40%	40%	40%	90%
Harry Potter 1	50%	60%	70%	70%
Ice Age 3	50%	40%	50%	70%
Jurassic World	50%	50%	50%	60%
Lilo & Stitch	50%	60%	40%	80%
Men in Black 3	50%	70%	70%	70%
Rampage	30%	30%	50%	80%
Shrek 2	40%	60%	30%	80%
Spider-Man	40%	60%	60%	60%
The Chronicles of Narnia 1	60%	70%	60%	70%
The Purge: Anarchy	60%	60%	60%	70%
Accuracy <sub>mean</sub>	47.27%	54.55%	52.73%	72.73%



**Fig. 4** Bar chart representing *Accuracy* provided by K-Means (K), BIRCH (B), LDA (L), and Word2vec (W) for 50 microgenres and 11 movies

outperforms the other models in 7 out of 11 movies, especially in the movie ‘Alexander’, and provides same results in the remaining ones.

Table 2 shows examples of microgenre assignments for the movies ‘Alexander’ and ‘Lilo & Stitch’, providing qualitative insights into the performance of the different approaches. The Word2vec-based approach demonstrated superior performance in terms of semantic similarity. For instance, in the case of the movie ‘Alexander’, the other approaches assigned incorrect microgenres, such as ‘christmas’ or ‘lawyer’, while the Word2vec-based approach only assigned ‘medieval’, which although incorrect, still bears some relevance to the historical period of the movie. Similarly, for the movie ‘Lilo & Stitch’, the other three models assigned incorrectly microgenres such as ‘psychological’, ‘zombie’, or ‘sex’, which are not appropriate for this animation movie for children. In contrast, the Word2vec-based approach missed only two out of ten microgenres and was the only one capable of assigning microgenres closely related to the movie plot, such as ‘alien’ and ‘family’.

In addition, Word2vec stands out as a completely automated model, as opposed to clustering and topic modeling techniques that require manual intervention to perform the matching process. Thanks to its semantic capabilities, Word2vec automatically associates each tag with its most related microgenres, removing the need for time-consuming manual operations. This automation not only improves workflow, but also provides greater accuracy in the matching process. Consequently, the forthcoming experiments will exclusively be conducted with this approach, affording the capability to work with all 666 microgenres.

### 3.4 Analysis of the effect of review tags

As a semantic model, Word2vec works well with tags that describe situations, characters, or scenarios related with the movie in a literal way. However, the reviews from customers contain many expressions used in an ironic or figurative sense. Table 3 shows some examples

**Table 2** Assignment of 10 out of 50 microgenres to the movies ‘Alexander’ and ‘Lilo & Stitch’ by K-Means, BIRCH, LDA, and Word2vec. Correct microgenres are in bold

K-Means	BIRCH	LDA	Word2vec
Alexander			
<b>history</b>	fantasy	<b>war</b>	<b>military</b>
<b>war</b>	<b>war</b>	<b>adventures</b>	<b>sex</b>
science fiction	<b>sex</b>	<b>drama</b>	<b>war</b>
suspense	<b>biopic</b>	mystery	<b>action</b>
animation	mystery	lawyer	<b>epic</b>
christmas	<b>adventures</b>	<b>action</b>	<b>history</b>
<b>adventures</b>	suspense	satire	<b>travel</b>
mystery	christmas	psychological	<b>romance</b>
<b>biopic</b>	monster	religion	medieval
accident	accident	teen	<b>drama</b>
Lilo & Stitch			
religion	<b>fantasy</b>	<b>children</b>	<b>animation</b>
<b>animation</b>	<b>monster</b>	sex	<b>animal</b>
<b>science fiction</b>	sex	psychological	<b>alien</b>
suspense	<b>animation</b>	drama	<b>comedy</b>
sex	romance	<b>adventures</b>	<b>children</b>
<b>animal</b>	<b>adventures</b>	military	<b>family</b>
drama	suspense	<b>mystery</b>	documentary
<b>space</b>	<b>space</b>	satire	superhero
<b>accident</b>	<b>comedy</b>	teen	<b>travel</b>
medieval	zombie	<b>comedy</b>	<b>fantasy</b>

of reviews for four of the considered movies. Figurative expressions like ‘stop beating around the bush’ or personal opinions like ‘Jurassic World is a big-budget indictment of corporate greed’ do not provide a useful contribution in this context (assignment of microgenres to movies). In addition, tags like ‘cheesy’ or ‘cigar’ are out of context, so they also introduce noise to the Word2vec-based approach.

**Table 3** Examples of movie reviews containing ironic or figurative expressions

Movie	Review
Alexander	It took me two nights (2½ hours total) to watch this - as I couldn’t stay awake the first night. OK, let’s stop beating around the bush. This is just a dog...
Jurassic World	If you squint really hard, you might find this to be good old dumb fun. But honestly, how long can you do that before getting a headache? So basically, Jurassic World is a big-budget indictment of corporate greed, jammed with product placement for Samsung and Mercedes-Benz and Beats by Dre and Coca-Cola
Rampage	I can understand why this film is getting low ratings, because it can be very cliched and cheesy...
The Purgue: Anarchy	...The first movie was close but no cigar down to rather lazy and...

In order to assess the influence of the reviews on the proposed approach, the whole collection (249,346 associations between tags and movies) and a subset of it that does not consider tags from reviews (109,664 associations) are used. Table 4 shows the results for the Word2vec-based approach in the two different scenarios.

When the Word2vec-based approach is applied to the complete movie collection, a mean accuracy of 78.18% is obtained. It is increased to 86.36% when the tags from reviews are not considered. This relevant improvement of 8.18% is attributable to the lack of the noise that is introduced by the customers' reviews in the complete collection. In addition, the absence of review tags provided better accuracy means in 7 out of 11 movies, whereas not improving results were obtained in any movie when the complete collection (including the review tags) was used. Therefore, our suggestion is not to consider the tags from reviews, since an improvement on accuracy is obtained. In the next experiment, the activation functions will be compared for the Word2vec-based approach by using the database that does not consider tags from reviews.

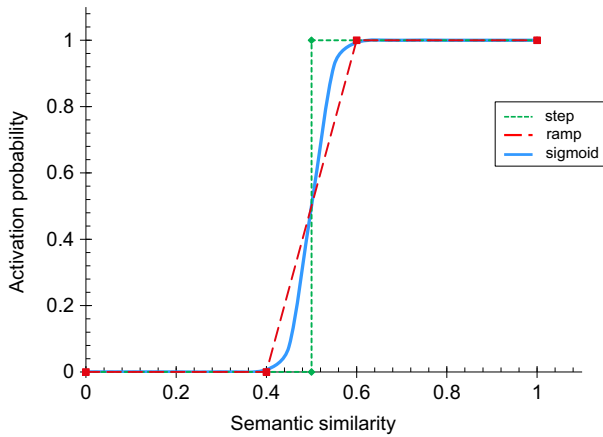
### 3.5 Comparison of activation functions

In previous experiments, the Word2vec-based approach used the activation function called binary step function, which assigns tags to those microgenres where the semantic similarity is greater or equal than a predefined threshold. This activation function is stricter, discarding more assignments between tags and microgenres. In this subsection, binary step function is compared with two other activation functions (ramp and sigmoid functions, see Section 2.4), in order to clearly analyze the differences among them.

Figure 5 graphically represents the best configuration for each activation function over the considered tag collection. Beyond a visual appreciation of their resemblance, these activation functions have significant differences in their slopes. The binary step function is the most restrictive function, showing a steep slope and being highly sensitive to changes. In contrast, ramp function has the least inclined slope, enabling a more gradual transition in output. Finally, sigmoid function, which has an intermediate inclined slope, provides a smooth transition as it moves toward the extremes of its activation range. This makes sigmoid function to be the most permissive activation function.

**Table 4** Accuracy and Accuracy<sub>mean</sub> provided by the Word2vec-based approach using the complete collection and without review tags for 666 microgenres and 11 movies

Movie	Accuracy <sub>complete</sub>	Accuracy <sub>NOreviews</sub>
Alexander	70%	70%
Harry Potter 1	70%	70%
Ice Age 3	80%	90%
Jurassic World	70%	80%
Lilo & Stitch	90%	100%
Men in Black 3	80%	90%
Rampage	90%	100%
Shrek 2	70%	90%
Spider-Man	70%	90%
The Chronicles of Narnia 1	90%	90%
The Purge: Anarchy	80%	80%
Accuracy <sub>mean</sub>	78.18%	86.36%



**Fig. 5** Graphical representation of the best configuration for binary step, ramp, and sigmoid functions

Table 5 shows the accuracy and linking rate metrics for the three activation functions. Binary step function, evaluated in the previous experiments (Section 3.4), provided a mean accuracy of 86.36%. In addition, its mean linking rate was 22. The ramp and sigmoid functions provided a mean accuracy of 87.27% and 86.36%, and a mean linking rate of 23.82 and 51.27, respectively.

The three activation functions provided very similar accuracies, being the ramp function the best alternative by only 0.91%. Since this marginal difference in mean accuracy does not conclusively state any activation function as the best one, a closer examination focusing on mean linking rate is considered. On the one hand, binary step and ramp functions provide closely similar mean linking rates. On the other hand, the sigmoid function outperforms the other activation functions by doubling the mean linking rate, while maintaining the matching quality (accuracy). This notable increment indicates a greater granularity in microgenre assignment. When movies have more tags related to different microgenres, it means they might fit into these categories more precisely or demonstrate greater diversity in their representation across many microgenres. This helps us to better categorize movies into categories and get a deeper understanding of what makes them unique.

## 4 Conclusions

The automatic assignment of microgenres to movies has been addressed as a relevant and current problem in media services. Three different techniques (clustering, topic modeling, and word embedding) have been considered to develop several matching approaches. Specifically, approaches based on K-Means, BIRCH, LDA, and Word2vec have been designed, implemented, and applied to a movie collection.

The results demonstrate that the Word2vec-based approach outperforms both clustering and topic models in terms of accuracy. This approach becomes the best one due to its semantic capabilities, model performance, and automation, making it a suitable choice for automatic microgenre assignment.

An analysis of the movie dataset revealed that excluding review tags improves the performance of the Word2vec-based approach by reducing noise introduced by these tags. This

**Table 5** Accuracies and linking rates provided by the Word2vec-based approach using binary step, ramp, and sigmoid functions for 666 microgenres and 11 movies

Movie	<i>Accuracy<sub>step</sub></i>	<i>Accuracy<sub>ramp</sub></i>	<i>Accuracy<sub>sigmoid</sub></i>
Alexander	70%	70%	80%
Harry Potter 1	70%	70%	70%
Ice Age 3	90%	100%	90%
Jurassic World	80%	90%	80%
Lilo & Stitch	100%	100%	100%
Men in Black 3	90%	80%	100%
Rampage	100%	100%	100%
Shrek 2	90%	80%	90%
Spider-Man	90%	90%	80%
The Chronicles of Narnia 1	90%	100%	80%
The Purge: Anarchy	80%	80%	80%
<i>Accuracy<sub>mean</sub></i>	86.36%	87.27%	86.36%
Movie	<i>Linking<sub>step</sub></i>	<i>Linking<sub>ramp</sub></i>	<i>Linking<sub>sigmoid</sub></i>
Alexander	23	26	66
Harry Potter 1	18	20	35
Ice Age 3	21	22	48
Jurassic World	19	20	62
Lilo & Stitch	29	29	57
Men in Black 3	16	19	32
Rampage	17	16	50
Shrek 2	24	29	54
Spider-Man	37	39	79
The Chronicles of Narnia 1	31	34	55
The Purge: Anarchy	7	8	26
<i>Linking<sub>mean</sub></i>	22	23.82	51.27

highlights the significance of considering the quality and relevance of tags in microgenre assignment tasks.

The optimization of tag assignment to microgenres was explored by investigating three activation functions: binary step, ramp, and sigmoid. Notably, these functions yielded similar accuracy results. Despite, the sigmoid function, with its higher mean linking rate, emerged as the more effective choice. By doubling the number of tags assigned to each movie while maintaining matching quality, the Word2vec-based approach utilizing the sigmoid activation function becomes a suitable approach for addressing the microgenre assignment problem.

As future research, Word2vec-based approach will be used to generate a more comprehensive taxonomy. This involves incorporating an additional classification level with a reduced number of genres to encapsulate the microgenres. Furthermore, expanding the application of this approach to a larger movie dataset, encompassing both movies and series episodes, will provide a more robust evaluation of its effectiveness.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11042-023-17442-y>.

**Acknowledgements** We are thankful to Optiva Media and Metadatol for providing the data necessary to conduct this study. This research has been supported by Ministry of Science, Innovation, and Universities – Spain and State Research Agency – Spain (Projects PID2019-107299GB-I00 and PID2021-122209OB-C32 funded by MCIN/AEI/10.13039/501100011033), Junta de Extremadura – Spain (Projects IDA3-19-0001-3, GR21017, and GR21057), and European Union (European Regional Development Fund).

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Data availability** Data have been included as electronic supplementary material.

## Declarations

**Conflict of interests** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Devine P, Blincoe K (2022) In: 2022 IEEE/ACM 1st international workshop on natural language-based software engineering (NLBSE), pp 1–8. <https://doi.org/10.1145/3528588.3528652>
2. Ullah I, Khusro S (2023) On the analysis and evaluation of information retrieval models for social book search. *Multimedia Tools Appl* 82(5):6431–6478. <https://doi.org/10.1007/s11042-022-13417-7>
3. Quintanilla E, Rawat Y, Sakryukin A, Shah M, Kankanhalli M (2020) Adversarial learning for personalized tag recommendation. *IEEE Trans Multimedia* 23:1083–1094. <https://doi.org/10.1109/TMM.2020.2992941>
4. Wang C, Yang X, Ding L (2021) Deep learning sentiment classification based on weak tagging information. *IEEE Access* 9:66509–66518. <https://doi.org/10.1109/ACCESS.2021.3077059>
5. Khan UA, Martínez-Del-Amor MA, Altowajri SM, Ahmed A, Rahman AU, Sama NU, Haseeb K, Islam N (2020) Movie tags prediction and segmentation using deep learning. *IEEE Access* 8:6071–6086. <https://doi.org/10.1109/ACCESS.2019.2963535>
6. Bizzocchi J (2020) Berlin remix—a computationally generative “city film” artwork. *Dig Stud/Le champ numérique* 10(1). <https://doi.org/10.16995/dscn.376>
7. Stevens AH, O'Donnell MC (2020) The microgenre: a quick look at small culture. Bloomsbury Academic, New York, NY. <https://doi.org/10.5040/9781501345845>
8. Wu C, Wang C, Zhou Y, Wu D, Chen M, Wang JH, Qin J (2020) Exploiting user reviews for automatic movie tagging. *Multimedia Tools Appl* 79(17):11399–11419. <https://doi.org/10.1007/s11042-019-08513-0>
9. Guehria S, Belleili H, Azizi N, Belhouari SB (2020) In: International conference on intelligent systems design and applications, Springer, pp 478–487. [https://doi.org/10.1007/978-3-030-71187-0\\_44](https://doi.org/10.1007/978-3-030-71187-0_44)
10. Yu Y, Lu Z, Li Y, Liu D (2021) ASTS: attention based spatio-temporal sequential framework for movie trailer genre classification. *Multimedia Tools Appl* 80(7):9749–9764. <https://doi.org/10.1007/s11042-020-10125-y>
11. Kundalia K, Patel Y, Shah M (2020) Multi-label movie genre detection from a movie poster using knowledge transfer learning. *Augment Human Res* 5(1):11. <https://doi.org/10.1007/s41133-019-0029-y>
12. Mangolin RB, Pereira RM, Britto AS, Silla CN, Feltrim V, Bertolini D, Costa YMG (2022) A multimodal approach for multi-label movie genre classification. *Multimedia Tools Appl* 81(14):19071–190966. <https://doi.org/10.1007/s11042-020-10086-2>
13. Nomoto T (2023) Keyword extraction: a modern perspective. *SN Comput Sci* 4(1):92. <https://doi.org/10.1007/s42979-022-01481-7>



14. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) In: Advances in neural information processing systems, pp 3111–3119
15. Ahmed MH, Tiun S, Omar N, Sani NS (2023) Short text clustering algorithms, application and challenges: a survey. *Appl Sci* 13(1):342. <https://doi.org/10.3390/app13010342>
16. Ahmed M, Seraj R, Islam SMS (2020) The k-means algorithm: a comprehensive survey and performance evaluation. *Electronics* 9(8):1295. <https://doi.org/10.3390/electronics9081295>
17. Ramadhani F, Zarlis M, Suwilo S (2020) In: IOP conference series: materials science and engineering, vol 725, p 012090. <https://doi.org/10.1088/1757-899X/725/1/012090>
18. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A (2023) Topic modeling algorithms and applications: a survey. *Inf Syst* 112:102131. <https://doi.org/10.1016/j.is.2022.102131>
19. Chauhan U, Shah A (2021) Topic modeling using latent Dirichlet allocation: a survey. *ACM Comput Surv* 54(7):145. <https://doi.org/10.1145/3462478>
20. Birunda SS, Devi RK (2021) Innovative data communication technologies and application, Springer, Singapore, pp 267–281. [https://doi.org/10.1007/978-981-15-9651-3\\_23](https://doi.org/10.1007/978-981-15-9651-3_23)
21. Wang B, Wang A, Chen F, Wang Y, Kuo CCJ (2019) Evaluating word embedding models: methods and experimental results. *APSIPA Trans Signal Inf Process* 8(1):e19. <https://doi.org/10.1017/ATSIP.2019.12>
22. Xia C, He T, Li W, Qin Z, Zou Z (2019) In: 2019 IEEE 19th international conference on software quality, reliability and security companion (QRS-C), pp 354–357. <https://doi.org/10.1109/QRS-C.2019.00072>
23. Zhang Y, Chen F, Zhang W, Zuo H, Yu F (2020) In: 2020 the 3rd international conference on big data and education, pp 37–42. <https://doi.org/10.1145/3396452.3396460>
24. Mikolov T, Chen K, Corrado G, Dean J (2013) In: International conference on learning representations 2013
25. McCallum AK (2021) MALLET: a machine learning for language toolkit. <https://mimno.github.io/Mallet> Accessed: 10 Sep 2023

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Carlos González-Santos<sup>1</sup> · Miguel A. Vega-Rodríguez<sup>2</sup>  ·  
Joaquín M. López-Muñoz<sup>3</sup> · Iñaki Martínez-Sarriegui<sup>3,4</sup> · Carlos J. Pérez<sup>1</sup>

Carlos González-Santos  
carlosgs@unex.es

Joaquín M. López-Muñoz  
joaquin.lopez@optimamedia.com

Iñaki Martínez-Sarriegui  
inaki.martinez@metadatol.com

Carlos J. Pérez  
carper@unex.es

<sup>1</sup> Departamento de Matemáticas, Universidad de Extremadura, Campus Universitario s/n, 10003 Cáceres, Spain

<sup>2</sup> Departamento de Tecnología de Computadores y Comunicaciones, Universidad de Extremadura, Campus Universitario s/n, 10003 Cáceres, Spain

<sup>3</sup> Research & Innovation Area, Consultora de Telecomunicaciones Optiva Media SL, Musgo 2 - Edificio Europa II, 28023 Madrid, Spain

<sup>4</sup> Research and Development Department, Metadatol SL, Santa Cristina 3 - Edificio Garaje 2.0, 10195 Cáceres, Spain